NRL Report 9189

Voice/Data Integration in Mobile Radio Networks: Overview and Future Research Directions

JEFFREY E. WIESELTHIER

Communication Systems Engineering Branch
Information Technology Division

ANTHONY EPHREMIDES

Locus, Inc.
Alexandria, Virginia

and

University of Maryland
College Park, Maryland

September 30, 1989

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a REPORT SECURITY CLASSIFICATION
UNCLASS

1b RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3 DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
NRL Report 9189

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Naval Research Laboratory

6b OFFICE SYMBOL *(If applicable)*
Code 5521

7a. NAME OF MONITORING ORGANIZATION

6c. ADDRESS *(City, State, and ZIP Code)*
Washington, DC 20375-5000

7b. ADDRESS *(City, State, and ZIP Code)*

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
Office of the Chief of Naval Research

8b. OFFICE SYMBOL *(If applicable)*

9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS *(City, State, and ZIP Code)*
Arlington, VA 22217

10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO.
61153N	RR021-05-42		DN480-557

11. TITLE *(Include Security Classification)*
Voice/Data Integration in Mobile Radio Networks: Overview and Future Research Directions

12. PERSONAL AUTHOR(S)
Wieselthier, J. E. and Ephremides, A.*

13a. TYPE OF REPORT

13b TIME COVERED
FROM ________ TO ______

14. DATE OF REPORT *(Year, Month, Day)*
1989 September 30

15 PAGE COUNT
84

16 SUPPLEMENTARY NOTATION
*A. Ephremides is with the University of Maryland and Locus, Inc.

17 COSATI CODES

FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Communications network
Voice/data integration
Radio network
Military communications
Discrete event dynamic system

19 ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This report gives the results of preliminary background investigations into the problem of voice/data integration in mobile radio networks. The nature of these networks creates a totally new environment, in which the problems of voice/data integration have scarcely been addressed.

We review the basics of the voice/data integration problem, and we discuss some of the approaches that have been developed in recent years. Few accurate models are available for optimization and control, and these have generally been limited to operation at a single network node. The problem of overall network operation and control is extremely complicated, even when issues related to mobile radio network operation are not addressed.

We outline our future research plans for this basic research task, which will stress the development of accurate analytical models for performance evaluation and control. Optimization models will be developed by using techniques such as Markovian decision process models. An important goal of this study is the development of a framework for overall network modeling. Toward this goal we plan to develop new network models based on the use of discrete event dynamic systems techniques.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT
□ UNCLASSIFIED/UNLIMITED □ SAME AS RPT □ DTIC USERS

21 ABSTRACT SECURITY CLASSIFICATION
UNCLASS

22a NAME OF RESPONSIBLE INDIVIDUAL
Jeffrey E. Wieselthier

22b TELEPHONE *(Include Area Code)*
(202) 767-3043

22c OFFICE SYMBOL
Code 5521

Previous editions are obsolete.

CONTENTS

VOICE/DATA INTEGRATION IN MOBILE RADIO NETWORKS:
Overview and Future Research Directions

1. INTRODUCTION

The traditional approach in communication network design has been to treat voice traffic completely separately from data traffic. This is because voice and data impose conflicting requirements on the networks that must support them. For example, voice communication requires that there be very low variability of the time delay, so that a continuous voice stream is presented to the user at the destination. In addition, low delay is generally desirable. In contrast, data traffic can often tolerate both long delay (except for certain time-critical messages) and high variability of delay. Also, data transmission requires very low probability of error to ensure the preservation of data integrity, while voice communication can tolerate considerably higher error rates as a result of the inherent redundancy of speech signals.

Many approaches for voice/data integration have been attempted during the past decade. Under the "boundary" method for multiplexing at a single node, sometimes known as the Slotted Envelope Network (SENET) scheme,* a fraction of the bandwidth of a node's outgoing link is allocated for voice and the remainder is available for data. Efforts in the commercial sector have concentrated on the "movable-boundary" method, under which the fractions are adaptively varied in response to traffic demands. But this work has emphasized the establishment of Integrated Services Digital Network (ISDN) standards, rather than the development of exact analytical models for such networks. Very little has been done to develop models that permit optimization. It has been necessary to use approximations to evaluate the performance of the movable boundary method, except in trivial cases.

Furthermore, analysis generally has been limited to the case of a single isolated node, although some results have recently been obtained for tandem configurations. In practice, most networks will have to support communication over multihop paths, where intermediate nodes are used as relays to connect platforms that are not within direct communication range of each other. The development of models that can trace performance over more than just a single link is an especially critical issue when voice is involved because the setup of a multihop voice call requires coordination with all nodes along the path, i.e., a real-time commitment of resources at each node. Generally, it has been necessary to use simulation to study such networks because analytical techniques are not available. In addition, the interrelationships between voice/data integration and other aspects of network design, such as voice compression, error control, channel access, routing, and congestion control have not been addressed adequately.

Manuscript approved December 16, 1988.

*Although the SENET scheme was originally introduced for use with a T1 carrier (as discussed in Section 7), the term SENET is often applied to any boundary scheme for voice/data integration.

Most studies of voice/data integration schemes have considered wireline networks or broadband satellite links, which are characterized by contention-free operation. Furthermore, their scope generally has been limited to the problem of multiplexing at a single node. In this study we are interested primarily in mobile radio networks, whose nature creates a totally new networking environment in which the problems of voice/data integration have scarcely been addressed.

It is primarily the "broadcast" property of radio networks that distinguishes them from wireline networks. This means that transmissions can generally be assumed to be heard by or interfere with all neighbors of a node, rather than only the one for whom the transmission is intended. Thus there is a much greater degree of interaction among network nodes than in wireline or satellite networks. Channel access schemes must be developed that reflect the specific needs of voice and data traffic, as well as the characteristics of the broadcast radio medium.

Another critical feature of mobile networks is their dynamic topology, which may result not only from actual platform motion, but also as a result of jamming, other-user interference, and changing propagation conditions. Network control schemes for routing and congestion control must be developed that are capable of adapting rapidly and robustly to changes in topology.

Few studies of voice/data integration in radio environments have addressed these issues, and virtually no attempts have been made to incorporate realistically the further complications that arise from considerations related to operation in a hostile Navy environment. Important issues that are pertinent to operation in such an environment include:

- the low data rate that can be supported by the communication channels in Navy networks (and the consequent impact on the feasibility of statistical multiplexing schemes that normally require high data-rate channels);
- security issues such as overhead and synchronization delay;
- the need to incorporate user and message precedences (e.g., the possible preemption of an ongoing call);
- the need to operate in hostile environments that are characterized by high levels of background noise, other-user interference, jamming, and fading, and the consequent use of spread spectrum signaling that leads naturally to the use of code division multiple access (CDMA) techniques.

Performance measures must be defined that consider the needs of both voice and data traffic and the nature of the radio channel. There will generally be a trade-off between voice performance measures (e.g., blocking probability, speech quality, and variability of delay) and data performance measures (delay, throughput, and error probability). In fact, as in any network, there will be trade-offs between the criteria specifically associated with voice and those specifically associated with data.

We have initiated a new study of voice/data integration methods for mobile military radio networks. This report contains an overview of the major issues that must be addressed when designing such networks and represents our understanding of this problem after a short period of preliminary investigation. We emphasize at the outset that our goal is to develop analytical models wherever possible, rather than to rely solely on simulation to evaluate system performance. Furthermore, we seek the development of control schemes for performance optimization, rather than simply the evaluation of ad hoc schemes. The use of Markovian decision process (MDP) models has recently resulted in some degree of success for single-node and tandem configurations. However, these techniques and other traditional methods have not been adequate for overall network modeling, and we expect to explore new methods that may prove to be more fruitful. In particular, we will investigate the application of discrete event dynamic system (DEDS) techniques to the design of communication networks. This class of approaches shows a great deal of promise, not only for the evaluation of system performance and the verification of system consistency, but also for providing a framework for system control and optimization.

In conclusion, many aspects of the voice/data integration problem require further basic research, even when considering only the case of wireline networks. When the further complications of the radio environment are considered, as well as the issues that relate specifically to Navy operation, it is clear that the study of this problem is still in its infancy.

Outline of the Report

Section 2 discusses the major aspects of the voice/data integration problem. We discuss the needs that must be satisfied by networks for these two very different classes of traffic, and we study their impact on network design and performance.

Section 3 discusses multiplexing and switching techniques for use in integrated networks. For most applications, either circuit switching or virtual circuit switching is best for voice, while the datagram mode of packet switching is best for bursty interactive data. To satisfy the needs of both types of traffic, a hybrid switching scheme is generally used for integrated networks, in which voice is circuit switched while data is packet switched. Section 4 discusses the characteristics of mobile military radio networks that distinguish them from wireline networks for which voice/data integration schemes are generally considered.

Section 5 discusses Integrated Services Digital Networks (ISDN), which have been developed primarily for commercial applications. Studies of ISDN have emphasized the development of standards, rather than the development of analytical models for performance evaluation or control.

Section 6 discusses speech interpolation methods, which are used to increase the number of calls that can be handled by a wideband communication link. We address the characteristics of military communication channels that distinguish them from the commercial ISDN environment. In particular, the low data rate that can be supported by these channels limits the speech interpolation advantage that can be achieved.

Section 7 discusses the problem of voice/data integration at a single node. In particular, we emphasize the movable boundary scheme, which is the most widely used approach for the dynamic allocation of a multiplexer's bandwidth to voice and data. We discuss several variations of this method and outline research that has appeared on this method in recent years.

Few substantive results have been obtained on network wide control of integrated voice/data networks. Section 8 reviews some of the more notable results that have been obtained to date.

Section 9 discusses the problem of channel access in integrated networks. We show how the different requirements associated with voice and data traffic affect channel-access decisions, and we discuss some possible approaches.

Section 10 discusses the problem of communication under time constraints. This is an important issue for voice communication, which must be delivered in nearly real time, as well as for certain types of tactical data traffic.

We believe that the modeling of communication networks as discrete event dynamic systems (DEDS) will provide a useful framework that will permit the development of methods for performance evaluation and control. Section 11 discusses techniques that have been used to model DEDS, as well as some possible applications of DEDS techniques to our problem.

Finally, in Section 12 we present a brief summary of the status of voice/data integration research in the overall research community, and we discuss our future research directions.

The study of optimization methods will be an important part of our studies. Appendix A presents a generic optimization approach that is applicable to the voice/data multiplexing problem at a single node, based on the formulation of a related MDP problem. We discuss three methodologies to approach this problem, namely dynamic programming, stochastic dominance, and linear programming. In Appendix B we demonstrate how almost any queueing control problem that can be formulated as an MDP can be converted into an equivalent linear program. In Appendix C we formulate the optimal control of a movable boundary scheme as an MDP, and we demonstrate how linear programming can be used to obtain the structure of the optimal policy. Extensions to tandem node configurations and the routing problem are also addressed.

2. GENERAL ISSUES—HOW VOICE DIFFERS FROM DATA

In communication network design, the traditional approach has been to treat voice completely separately from data. The reason for this is that voice and data impose conflicting requirements on the networks that must support them. For example, voice communication requires that there be very low variability of the time delay, so that a continuous voice stream is presented to the user at the destination. Also, low delay is required to support normal interactive speech. In contrast, data traffic can usually tolerate relatively long delay (except for certain time-critical messages) and high variability of delay. Thus, in this regard voice communication imposes a more severe requirement on network operation. On the other hand, voice traffic can tolerate a considerably higher error rate than data traffic because of the inherent redundancy in human speech. From this perspective, data communication imposes the more severe requirement.

Performance measures must be defined that consider the needs of both voice and data traffic and the nature of the radio channel. There will generally be trade-offs between voice performance measures (e.g., blocking probability, speech quality, and variability of delay) and data performance measures (delay, throughput, error probability, and packet loss resulting from buffer overflow). In fact, there will be trade-offs among the criteria specifically associated with voice; similarly, there will be trade-offs among the criteria specifically associated with data. Many of the issues associated with voice/data integration are addressed by Pickens and Hanson (1985). Also of particular interest is a special issue of the IEEE Journal on Selected Areas in Communications (Decina and Vlack, ed. 1983), which is dedicated to this important topic. Specifically, in this issue Gruber and Le (1983) address performance requirements for integrated networks. Several other papers from this issue are also referenced in this report.

Voice/data integration schemes have been developed for wireline networks and broadband satellite links. However, the nature of radio channels creates a totally new networking environment in which the problems of voice/data integration have scarcely been addressed. As we discuss later in this report, radio networks are characterized by a greater degree of interaction among network nodes than are wireline networks, thus resulting in more complicated network design and analysis. In addition, military considerations have an important impact on network designs. Of crucial importance are survivability in stressed environments and security issues.

Voice traffic consists of "calls," a terminology that suggests the continuing need for use of network resources throughout the duration of the exchange. Data traffic, on the other hand, is characterized by "messages," which often consist of short transmissions that may be divided into smaller units known as packets. In this report, the terminology of call and message always refers to voice and data traffic respectively. In this section we address some of the major differences between voice and data and their impact on integrated network design.

Data Traffic

The primary performance criteria used to evaluate data networks are usually throughput and delay. Implicit in the throughput requirement is that the data be delivered with a low bit-error rate (BER); e.g., a BER of 10^{-3} or 10^{-5} is often specified. When appropriate error-detection codes are used, automatic repeat request (ARQ) schemes can be implemented in which packets with errors are retransmitted, thus resulting in nearly error-free performance.*

The magnitude of delay that can be tolerated by data depends greatly on the type of traffic that must be supported and typically may range from seconds to hours. Whether ARQ schemes can or cannot be used depends largely on whether or not data packets can tolerate the delay associated with retransmission. Also, ARQ cannot be used in applications where a reliable feedback channel does not exist between the destination and the source, e.g., in applications where the destination is radio-silent or jammed, or if a contention-free channel for feedback is not available. When ARQ is not feasible, error-correction codes must be used that are sufficiently powerful to enable packets to be delivered correctly at their first attempt with the specified probability. The delay criterion is especially important for bursty traffic, which may consist of one or more packets. For multipacket messages, it is not critical that packets be delivered with uniform delay, or even that they be delivered in the sequence in which they were generated; it is essential that enough information be transmitted with each packet to permit resequencing at the destination. Delay is less important for bulk traffic, which consists of long file transfers. Such traffic is often handled with a lower precedence rating and is transmitted only when the network is not busy with higher precedence traffic.

Data traffic can usually be buffered at the node at which it is generated, as well as at relay nodes along a store-and-forward path in a multihop network. Since, in practice, data buffers are finite, the probability of packet loss as a result of buffer overflow can be an important performance measure in times of congestion. This problem is often ignored because most analyses assume the availability of infinite buffers.

*Clearly, the need to retransmit packets results in reduced throughput.

Many models for data traffic have been considered. It is often assumed that data traffic is bursty and consists of single fixed-length packets that are characterized by a Poisson arrival process. The use of fixed-length packets makes it easy to divide the time axis into fixed-length slots, which serve as the framework for a variety of channel-access schemes. Many channel-access schemes have been developed that reflect the requirements of bursty traffic. If we consider the case in which traffic is regular, e.g., if each user may generate a single-packet message periodically, TDMA is appropriate. When long data transfers are considered (which may be generated in a bursty manner), the approaches developed for bursty single-packet traffic are no longer appropriate; schemes that reflect the continuing need for the use of the channel resource should be used. In such cases the characteristics of data traffic are somewhat similar to those of voice, although without the real-time delivery requirement. We may also consider the intermediate situation in which data messages consist of several packets (e.g., geometrically distributed). In general, the development of schemes for channel access and overall network control must reflect the traffic requirements of the particular scenario for which the network is being designed.

Voice Traffic

The primary characteristic of interactive voice traffic is that it must be delivered as a steady stream in real time.* Usually, the delays that are associated with ARQ schemes would not be tolerable. However, because of the inherent redundancy in human speech, relatively high error rates can be tolerated, and error control coding requirements are not very stringent. The delays experienced in voice networks may be broken down into several components. For example, in packet-switched multihop systems the components of delay that must be considered include packetization delay, the queueing delay at the intermediate nodes along the route, transmission delay, propagation delay, and delay at the receiving end.

It is difficult to develop an absolute requirement for speech delay, i.e., to determine the delay that would result in unacceptable speech performance. The difficulty stems primarily from the highly subjective judgment of speech quality. Few people observe quality degradation in interactive speech when delays are less than about 300 ms (Gold 1977; Gitman and Frank, 1978). When delays are larger (between 300 ms and 1.5 s), speakers notice the difference and change their speaking patterns; speech is then characterized by longer pause intervals and overlapping talk. Significantly larger delays may be tolerable in military environments, in which poor speech quality has long been accepted as a fact of life. particularly when the alternative to speech with long delays may be no speech at all.

In circuit-switched networks, delay is fixed because a fixed path is used and because transmission resources are allocated at each node along the path with no queueing delay. However, in packet-switched networks speech can experience significant variation in delay from one packet to the next. For example, packets may take different routes and possibly arrive out of sequence. Even if a fixed path is used, queueing delays may vary from packet to packet because of traffic fluctuations. The variability of the delay experienced by the individual segments of speech transfer is a critical issue in voice networks. This is because the receiving node must provide a continuous stream of bits to the vocoder to reproduce the speech waveform at the destination. If a packet is not available when it is needed, the information contained in it is lost; the receiver may either generate noise or silence (if packets are very small, some form of interpolation may be used), and degraded speech quality may result. The effects of speech loss can be quite severe if critical words are lost (e.g., "not"), thereby

*Here we are considering only real-time, interactive speech. One could also consider stored forms of speech ("voice mail"), which would be similar in delay requirements to bulk data transfer, with, however, the lower fidelity requirements of voice traffic.

changing the meaning of the phrase completely. For this reason it is necessary to preserve continuity over the duration of a talkspurt, which should be long enough to contain about one phrase or sentence (Gruber 1981). The effects of the variance in arrival times can be reduced by delaying packets in the receiver's buffer to permit smoothing of the received waveform. This, of course, adds to the delay of the voice stream and requires additional hardware (i.e., buffers) at the receiver.

Blocking probability is also an important performance criterion for voice traffic. In most applications (e.g., telephone), a call is not accepted by the network if transmission resources are not immediately available. In contrast, data traffic can usually be buffered and await transmission.

The inherent quality of the speech waveform (assuming delay and delay variance are not problems) is directly related to the data rate used to encode the digital speech waveform. For example, the standard for voice communication in commercial digital networks is a 64 kb/s waveform. Recent efforts to achieve compression have produced linear predictive coder (LPC) encoded voice at 2.4 kb/s that provides acceptable performance for military applications; recently NRL has developed an 800 b/s LPC voice processor that provides performance that is only slightly worse than that of the 2400 b/s LPC processor (Kang and Fransen 1985). Use of lower data rates to encode voice permits the sharing of channel bandwidth by several voice streams. Section 6 describes the trade-offs between network throughput and voice quality resulting from the use of lower rate voice.

Voice traffic is characterized, in part, in terms of the arrival rate (i.e., start of new conversations), the distribution of the length of the conversation, and the frequency with which the two talkers exchange their roles of speaking and listening. For example, it is often assumed that the lengths of both active and inactive periods are exponentially distributed. Daigle and Langford (1986) considered three queueing models for packet speech: a semi-Markov process model, a continuous-time Markov chain model, and a uniform arrival and service model. The suitability of these models depends on the parameters of the particular system being considered. Accurate models are needed to characterize the environment for which the particular networking scheme is being developed.

Speech traffic normally requires a two-way path between source and destination (possibly traversing several intermediate nodes) that must be maintained for the duration of the call. Section 6 discusses the properties of speech traffic that make speech interpolation schemes possible, thereby improving the efficiency of speech communication systems.

3. SWITCHING TECHNIQUES FOR INTEGRATED NETWORKS

Switching is the mechanism by which a network's resources (nodal processors and links) are allocated to the communication functions supported by the network. The switching function permits communication among a network's members without requiring the construction of permanent dedicated (multihop) links connecting all possible pairs of users. The switching mechanism is implemented in conjunction with a multiplexing technique, which permits the combination of data or voice from multiple users on the same communication link. In this section, we first provide a brief discussion of multiplexing techniques. We then discuss switching techniques, and we comment on their suitability for use in integrated networks.

Multiplexing Methods

Multiplexing permits the combining of data or voice from several senders on the same communication link. The suitability of a multiplexing scheme for a particular application depends on the type of traffic that must be supported and on the type of channels that are available to support this traffic. Our discussion in this subsection is brief and serves as an introduction to the subsequent material on switching.

Frequency-division multiplexing (FDM) is the traditional method of sharing channel bandwidth among a number of users. Under this method, a wideband channel* is divided into a number of smaller-bandwidth channels, each of which is capable of supporting the demands of one user. This subdivision of the channel is generally fixed; each channel is assigned to a pair of users for the duration of the all. Analog filters are used to separate the traffic streams corresponding to the different users.

A fixed allocation scheme can also be implemented in the time domain by using time-division multiplexing (TDM). Under TDM each user is granted the full bandwidth of the wideband channel on a periodic basis, i.e., once per fixed-length time frame. Varying communication requirements can be accommodated by assigning different-length time slots to the users as needed.

Operation in the time domain permits degrees of freedom that are not available in the frequency domain. For example, greater efficiency can be achieved when traffic is bursty by using schemes that allocate the channel on a dynamically assigned basis. With statistical-time-division multiplexing (STDM) (Chu 1969), the size of the slot apportioned to each user varies dynamically, based on traffic requirements. Alternatively, STDM schemes can be considered under which the time slots are of fixed duration but where a time-varying set of users transmits in each frame, based on current user demands. In either case, a composite packet is constructed that contains contributions from some or all of the users at the node, as well as the addressing and other overhead information. At periods of high congestion, it may not be possible to accommodate the traffic of all users in the frame, resulting in increased buffering requirements or lost data. Thus, STDM permits a greater number of users to share a wideband channel than is possible under TDM. However, under STDM the users are not guaranteed a time slot in every frame.

Under packet multiplexing the traffic from each source is individually packetized and contains its own header and error-detection information. Like STDM, packet multiplexing adapts dynamically to user demands, but it is more flexible because packets can be transmitted without waiting for the construction of the composite packet. However, overhead is higher than under STDM because of the need to provide complete header information with every packet.

Comparison of Switching Techniques

The three basic forms of switching used in networks are circuit switching, message switching, and packet switching. This section briefly discusses these mechanisms and some of their variations and comments on their suitability for use in integrated networks.

Circuit Switching

Under circuit switching (which is commonly used in telephone systems) a path (which generally passes through several nodes) is first established between the source and destination nodes before any information is transmitted. The resources associated with this circuit are then dedicated exclusively to

*In this discussion, we use the term wideband channel to refer to any channel that can support several baseband channels.

that particular call or transaction. The primary advantage of circuit switching is that virtually no overhead is incurred once the circuit is set up. The disadvantages of circuit switching are the initial delay experienced in setting up the circuit (and reestablishing a new circuit in the event of topological change) and the fact that the resources associated with the circuit cannot be used for any other purpose, even if the users of the circuit are temporarily idle. Circuit switching is best suited for long, continuous transactions, such as voice calls or bulk data transfers.

The development of faster switches will permit the use of fast circuit switching (Harrington 1980), under which paths can be established much more quickly, thus resulting in efficient operation even for relatively short transactions. However, such schemes are not feasible for radio networks because the major components of delay in radio networks are related to channel access, particularly in multihop applications, rather than to the speed of the switching elements.

Circuit switching can be implemented in conjunction with either FDMA or TDMA multiplexing schemes.

Message Switching

Under message switching, a message is stored and forwarded from node to node without following an assigned physical connection between source and destination. Messages are processed as complete units, and are thus, in general, of variable length. A complete message must be received at a node before it is forwarded.

Packet Switching

Packet switching is similar to message switching, but messages are divided into fixed-length blocks, which are individually sent through the network. Thus several packets corresponding to a single message can be in transmission simultaneously, often resulting in considerable reduction in delay as compared to message switching. This property is referred to as the "pipeline" effect. There are two distinct modes of packet switching, the datagram mode and the virtual circuit mode.

Under the datagram mode, each packet is processed independently by the network and must contain complete header information, including the complete address of the origin and destination.[*] Thus the overhead of the datagram mode can be considerably greater than that of message switching, under which the header information has to be transmitted only once per complete message. Multipacket messages must be pieced together at the destination from packets that may have been received out of order, thus generating the need for resequencing protocols.

Under virtual circuit switching, a logical channel (which generally uses a fixed path[†]) is established for the user pair. Once this is done, the identification (ID) number associated with the virtual circuit is transmitted instead of complete header information, resulting in considerable saving in overhead. In addition, this mode provides for the sequencing of packets within each session, eliminating the need for a resequencing protocol at the destination. There are two basic types of virtual circuit connections. First, virtual circuits may be set up on a temporary basis in response to a request by one of the members of the user pair, and then shut down when the call is complete; this is the switched virtual circuit configuration. Alternatively, a permanent virtual circuit may be assigned to a user pair.

[*] In military applications, security headers may be extremely long.

[†] It would be possible to establish virtual circuits that use varying paths, but to do so would be considerably more difficult

The number of virtual circuits that can be supported simultaneously by a network depends on the buffer space available at each node to store the routing information associated with each of the virtual circuits. Note that virtual circuits must be shut down and reconstructed whenever their operation is disrupted by topological changes, which may occur frequently in mobile radio networks (thus precluding the setup of truly permanent virtual circuits in many applications). Note that, unlike true circuit switching, virtual circuit switching does not guarantee a fixed-delay path through the network, even though a fixed path may be followed. The delay is variable because network resources are not dedicated to the exclusive use of a circuit, a situation analogous to STDM discussed earlier. If a virtual circuit cannot be supported during any particular frame (because all time slots have been assigned to other users), one approach is to store the untransmitted packets in a buffer. Thus variable queueing delays may be experienced, particularly when the network is heavily congested. Alternatively, packets may simply be dropped to ensure that the packets that are actually transmitted do, in fact, experience uniform delay. This statistical nature of virtual circuit switching is a distinct disadvantage for voice traffic, as compared with true circuit switching under which channel resources are guaranteed.

The virtual circuit mode provides two main advantages over the datagram mode. In addition to the reduction of overhead, the virtual circuit mode provides improved controllability of data flow, because the flow on individual virtual circuits can be controlled. For example, the flow on one virtual circuit (on all of its links) can be stopped temporarily until congestion is reduced. No such selective control can be achieved in datagram networks. Similarly, a virtual call would be accepted only if sufficient bandwidth resources were available within the network. A third advantage of virtual circuit switching is that it results in loop-free operation, since packets follow a fixed path through the network. In contrast, in the datagram mode packets are treated individually and looping may occur.

Despite its advantages, virtual circuit switching does exhibit some disadvantages compared to the datagram mode of operation. For example, the call setup phase for nonpermanent connections results in delays that may not be acceptable, particularly when transmissions are very short and when timing is critical. Also, increased memory is required at the nodes to store the routing information associated with the virtual circuits that are being supported, as well as to store the traffic associated with each of the virtual circuits. Furthermore, the restriction to the use of fixed paths prevents the exploitation of increased efficiency that may sometimes be obtained through dynamic routing mechanisms. In summary, virtual circuit switching can be characterized by the following features (Gerla 1985):

- virtual circuit setup prior to transmission;
- fixed path (established at call setup);
- dynamic sharing of bandwidth on the trunks along the path;
- sequencing of user data within each virtual circuit;
- selective flow control; and
- identification of user data within the network by means of an ID number rather than full source/destination address.

Gerla concluded that virtual circuit switching is more cost-effective than the datagram approach for applications where the following conditions are satisfied:

- An initial call setup delay of a few seconds is acceptable.
- The duration of the session is considerably longer than the call setup time.
- The average data flow rate within the session is sufficiently high to justify the dedication of virtual circuit resources (context blocks, buffers, etc.) along the path for the entire duration of the session.

Some of the traffic that will be carried by integrated military radio networks will meet these criteria, but certainly other traffic will not. In some applications, the use of a datagram mode of operation may permit a higher degree of survivability and efficiency For example, the need to adapt to frequent topological changes may make the use of virtual circuit switching difficult, because virtual circuits will have to be shut down and then reestablished in response to such changes. In contrast, datagram traffic can be supported during topological changes, although the derivation of dynamic routing procedures for such applications remains an area of investigation. Also, datagram operation may ultimately permit more efficient use of the channel, although at the expense of more complicated control mechanisms that have not yet been developed.

Switching Methods for Voice Communication

The traditional method for voice communication is, of course, circuit switching, as in the telephone network. We have noted that virtual circuit switching (which uses packet switching as the underlying data transport mechanism) can provide a service that appears to be very similar to that of true circuit switching. The feasibility of the use of virtual circuit switching has been demonstrated experimentally by using the ARPANET and SATNET (Weinstein and Forgie 1983). These studies concluded that packet communication techniques that use virtual circuit switching are, in fact, practical for real-time speech communication. This study did not conclude that packet-based virtual circuit schemes were preferable to circuit switching. However, it did conclude that if a packet data communication network is already available, it is more attractive economically to add a speech service to this network, rather than to provide an additional, completely separate speech service. In an earlier study, Coviello (1979) presented a comparative discussion of circuit-switched and packet-switched voice; he considered four schemes that ranged from a datagram mode of operation to a virtual-circuit approach and reached similar conclusions.

The requirements of real-time voice traffic (i.e., low delay combined with low delay variance) might suggest that either circuit switching or virtual circuit switching is the method of choice. This is not necessarily true, however. For example, the dynamic nature of mobile networks may render the implementation of either of these methods difficult or impossible, because communication paths may not be stable over the duration of a call. Also, the communication requirements may be quite different in military scenarios than in the commercial world. For example, delays of several seconds may have to be considered acceptable if the alternative is not to receive the call at all. This relaxation of delivery times makes the need for real-time connections less important than the reliability of the connections. True packet switching may provide the only means to deliver traffic (either voice or data) reliably over networks with rapidly changing topology. The control schemes needed to do so do not exist yet, however.

Very little has been accomplished in terms of developing speech communication methods for packet radio networks with mobile nodes. Shacham et al. (1983) discuss a networking scheme for such networks. Experiments have demonstrated that a ''duct-routing'' scheme can provide acceptable performance; this scheme is discussed in Section 8.

At this point it is worth mentioning that a variety of techniques have been developed to multiplex several voice calls over the same communication channel. Section 6 discusses time-assigned speech interpolation (TASI), speech predictive encoded communications (SPEC), digital speech interpolation (DSI), and multirate voice coding. These methods may be used in conjunction with hybrid switching techniques that permit the integration of voice with data.

Hybrid Switching Techniques

Section 2 notes that the different types of traffic that integrated networks must support are characterized by very different delivery requirements. For example, voice traffic requires real-time delivery (i.e., low delay combined with low delay variance) but not necessarily very low error rate. Data traffic requires extremely low error rate and may require low delay, but generally is not sensitive to variance in the delay. To accommodate the widely varying requirements of the different types of traffic that many networks must support (voice, interactive data, and file transfer), hybrid switching concepts have been proposed that combine the advantages of the various switching mechanisms that have been discussed in this section.

What is needed is a switching technique that can provide service similar to circuit switching (or virtual circuit switching) for voice traffic (or any other traffic that requires real-time delivery) and packet-switching for interactive data traffic. The elements of circuit and packet switching can be combined through the use of multiplexing methods that permit the sharing of channel capacity among a number of voice calls and data packets. Section 7 discusses the movable boundary scheme (sometimes referred to as SENET) and other methods of combining voice and data traffic.

Conclusions on Switching Techniques

In this section we have discussed the principal methods for switching, and we have noted the applications for which each is most suitable. For example, circuit switching is appropriate when communication requirements are uniform for relatively long periods of time. Virtual circuit switching (actually a form of packet switching) functions well when voice or data transfers are of relatively long duration but exhibit some degree of burstiness. The datagram mode of packet switching tends to be best for data traffic that consists of a small number of packets. Hybrid switching appears to be the best approach to combine these very different types of network traffic; this approach has been, in fact, the one most studied.

However, it is not possible at this point to draw any definite conclusions on the switching mode that is most suitable for integrated military radio networks. For example, the need to adapt to frequent topological changes (caused by platform mobility, changing propagation conditions, or changing interference conditions) may make the implementation of virtual circuit switching difficult. Packet-switched schemes may have to be developed that do, in fact, adapt to such dynamic environments. Continued study is needed to determine the best switching techniques for mobile military networks.

4. THE MOBILE MILITARY RADIO ENVIRONMENT

Network operation in a mobile radio network environment raises many issues that are not encountered in the design of networks for wireline applications or broadband satellite links. In either of these cases, contention-free links connect the network nodes. However, in radio networks the problem of channel access must be addressed because the signals of different users may interfere with each other. In fact, the nature of radio channels creates a totally new networking environment where the problems of voice/data integration have scarcely been addressed—an environment in which there is a considerably greater degree of interaction among network nodes than in wireline networks, as we now discuss.

It is primarily the "broadcast" property of radio networks that distinguishes them from wirelin^e^ networks. By this we mean that transmissions can generally be assumed to be heard by or interfere with all neighbors of a node, rather than by only the one for whom the transmission is intended. Thus the problem of channel access enters the picture. If contention-based channel access procedures are used, provision must be made for the retransmission of unsuccessful packets. This may be a considerable fraction of the packets transmitted over the channel (more than half of the packets transmitted may be unsuccessful in heavily congested systems). Analyses must reflect the lack of independence of operation at network nodes that results from the fact that all packets involved in a collision are normally assumed to be received incorrectly. In contrast, although packet errors may occur as a result of interference (e.g., background noise or jamming), in contention-free applications they are normally independent and relatively infrequent events. Note that contention-free operation can be achieved not only by scheduled TDMA transmissions but also by demand-assignment (reservation) schemes. One reasonable approach may be the use of scheduled transmissions to implement a reservation channel, which is used to establish contention-free channel allocations for the duration of a voice call. Section 9 addresses the problem of channel access in integrated networks.

Dynamic topology is another important feature of mobile networks. This may result not only from actual platform motion but also as a result of jamming, other-user interference, and changing propagation conditions. Networking schemes must be developed that are capable of adapting rapidly and robustly to changes in topology. In particular, it may not be possible to maintain circuit-switched or virtual-circuit-switched communication paths in rapidly changing environments. This will impact heavily on voice traffic, which normally relies on one of these switching modes. Thus it may be necessary to develop new datagram-type approaches that can support voice and data applications in such environments.

Few studies of voice/data integration in radio environments have been made that address these issues, and few attempts have been made to incorporate realistically the further complications that arise from considerations related to operation in a hostile Navy environment. A critical issue is the data rate that can be supported by the communication channels in the network. Since the data rates supported by military radio communication channels are generally much lower than those in commercial applications, it is not possible to multiplex as many signals onto the same waveform. Therefore, it is not possible to achieve the efficiency provided by the statistical multiplexing schemes that are often used in high bandwidth channels (see Section 6). Such schemes exploit the law of large numbers, which is not applicable when only a few users can transmit simultaneously by using the same waveform.

The need to incorporate user and message precedences (e.g., the preemption of an ongoing call) is critical in many military applications. Also, different performance measures must be considered. These include the delivery of time-critical messages (e.g., the probability of message delivery within time constraints), message delay, and channel throughput.

Security issues must be addressed, especially the impact of overhead and synchronization delay. For example, packet-switched operation would not be feasible if complete crypto header information must be transmitted with every packet. These issues are especially important in multihop operation, where synchronization delays may be accrued at every hop. Schemes must be developed that require a minimum amount of overhead.

Finally, Navy networks must operate in hostile environments that are characterized by high levels of background noise, other-user interference, jamming, and fading. The use of spread-spectrum signaling (e.g., frequency hopping), necessitated by antijam considerations, leads naturally to the use of code division multiple access (CDMA) techniques that permit several signals to share a wideband channel simultaneously. Multiple access concepts must, in fact, be redefined to take advantage of the flexibilities offered by CDMA operation (Wieselthier 1988). New performance models must be developed for CDMA channels that support voice/data integration. These models must reflect the different characteristics of the two types of network traffic and their complex interrelationships. For example, more powerful forward-error-control codes may be used for data than for voice. Also, the multiple-user capabilities of channels supporting data and voice simultaneously will have to be evaluated. Recently, Geraniotis and Gluck (1987) evaluated the performance of coded frequency-hopped systems in complex hostile environments. Similar models for channels that support two classes of traffic with different performance requirements have not yet been developed. Such models could facilitate the development of new types of flow-control schemes that reflect the need of the network to support both types of traffic. It may be desirable to divide the time axis into alternating intervals that support either voice or data traffic but not both simultaneously. In this way the ability of voice signals to tolerate higher interference levels will not impact adversely on data.

We conclude by noting that many aspects of the voice/data integration problem exist that require further basic research, even when considering only the case of wireline networks. In particular, major issues include dynamic optimization techniques and multihop operation, which we discuss in Sections 7 and 8. When the further complications of the radio environment are considered (such as channel access and dynamic topology), as well as the issues that relate specifically to Navy operation, it is clear that the study of integrated mobile military networks is still in its infancy, and that fundamental research must be carried out before conclusive design guidelines for practical systems can be formulated. Integrated switching models developed for commercial applications will be the starting point for our research, but it will be necessary to develop and analyze schemes specifically for military radio applications.

5. INTEGRATED SERVICES DIGITAL NETWORKS

Data communication (file transfers, electronic mail, facsimile, remote sensing, interactive or distributed computation, etc.) has experienced and continues to experience a rapid growth. This growth has created a pressing need for additional communication (i.e., switching and transmission) resources with increasingly enhanced features. To meet these enhanced needs, the concept of an Integrated Services Digital Network (ISDN) has been proposed (Stallings 1985; Decina et al. 1986a,b; Decina and Roveri 1987; Ronayne 1988). This section discusses efforts toward implementing the ISDN in the commercial sector, which represents most of the ISDN activity. We note at the outset that most of these efforts have emphasized the establishment of standards, rather than the development of analytical models for system performance. We also discuss potential military applications of ISDN, including issues specifically related to voice/data integration in mobile radio networks.

ISDN in Commercial Environments

The use of the existing telephone network, which represents a multibillion dollar investment, seems to be a sensible approach toward implementing data communications for a variety of reasons. First, the addition of new subscribers to the existing telephone network means increased utilization of its resources and, therefore, cost reduction resulting from more efficient network use. Second, from the users' point of view, a single, common network for both telephone and data communications (in other words, an integrated services network) provides uniformity of services. For example, separate sources or multifunctional terminals (e.g., voice, facsimile, data, videotex sets) generating traffic with differing characteristics and requirements would require a single interface to connect with a single network that supports all types of traffic.

An ISDN can be loosely described as a digital network that supports users with widely differing characteristics and requirements. Even though the telephone network is not yet fully digital, it is gradually migrating toward the envisioned ISDN network environment that will eventually use all-digital switching and transmission facilities. More specifically, we can characterize an ISDN by its three main features (Decina and Roveri 1987):

- end-to-end digital connectivity,
- multiservice capability (voice, data, video), and
- standard terminal interfaces.

The term ISDN is sometimes used to refer to any network that supports voice and data traffic. However, in this report we use the term to refer specifically to the integrated commercial networks that are currently under development and that conform to the standards that have been established for such networks.

ISDN has two fundamental roles. One is to provide fully digital interconnectivity for sophisticated voice and data applications. The other is to support and advance the distributed processing capability in a communications environment.

The fundamental objective of the ISDN from its inception was that it be based on the telephone network—this requires that most existing copper loops be able to carry ISDN traffic by simply adding electronics at the end. This led to the classic trade-off: future technology vs near-term availability. Future trends include all-digital transmission facilities (packetized voice on the existing networks or design of completely new ones) and integration of subscribers requiring bandwidth considerably higher than that currently offered by the telephone network (e.g., teleconferencing and video transmission).

ISDN could be implemented as a single, all-encompassing digital network with one integrated transport and switching fabric. The problem with this approach is the impracticality of replacing all of today's multibillion-dollar communications infrastructure with this single integrated digital network. The conceptual view that has unlocked the door to achieving ISDN is the realization that from the viewpoint of the end user, access to a series of otherwise independent special-purpose digital networks (e.g., circuit, private line, packet, wideband) is an effective first step. That is, as long as a user has this integrated network access, the network itself can be implemented in a variety of ways, all of which would be transparent to the user. Multiple, independent overlay networks (of which three are illustrated in Fig. 1) are linked together at a common interface point, which provides integrated access to the end user. If this interface is defined and standardized, network providers and end users can implement and evolve their respective subnetworks independently and still ensure compatibility. To the end user, the particular realization of the network is immaterial since the network capabilities are defined through the interface. Thus a focus on understanding ISDN becomes a focus on understanding ISDN access.

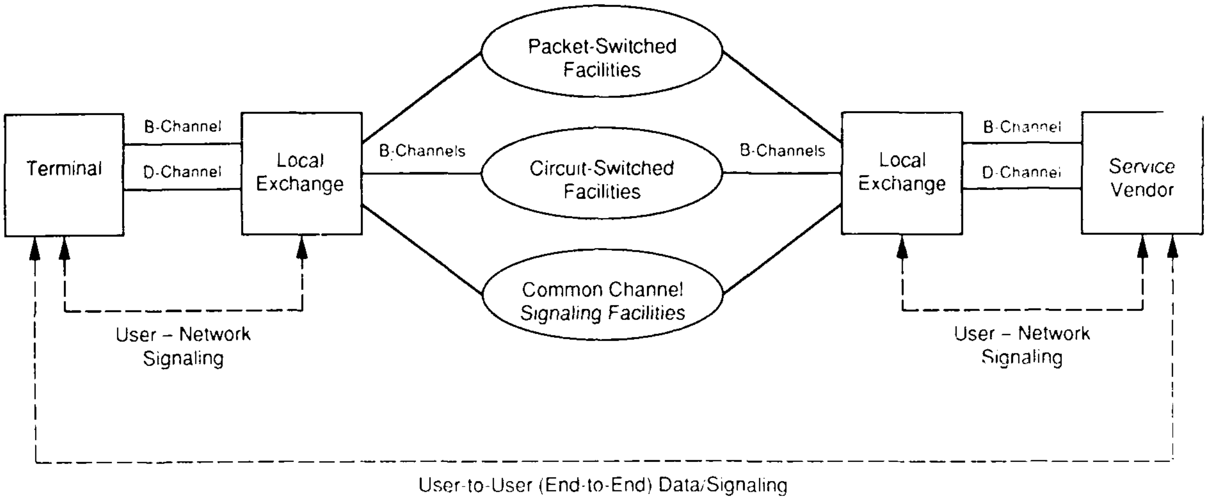

Fig. 1 — Basic ISDN architecture

ISDN access defines a completely digital interface subdivided into channels. These channels are of two general types: B channels, which provide a 64 kb/s service, are used only for customer information (voice, data, video); D channels, which provide a 16 kb/s service, are used for sending the signaling and control information that manages the information-carrying channels. Thus the signaling is "out-of-band" rather than "in-band." This approach has two main benefits: all the capacity in the information-bearing channels is available for customer use, and the special signaling channel allows for robust, distributed processing across the ISDN interface.

The International Telegraph and Telephone Consultative Committee (CCITT), an international agency of which the United States is a member, has been the principal forum for defining the technical details of ISDN. The committee has selected the fundamental digital rate building block as 64 kb/s. Furthermore, the committee has defined two major interfaces, the basic-rate interface (BRI) and the primary-rate interface (PRI). BRI is intended to serve information sources or sinks of relatively small capacity such as terminals. PRI is intended for large-capacity vehicles such as private branch exchanges (PBXs). Both have a similar structure—one D channel and a number of B channels. The basic-rate arrangement is 2B + D. The structure for primary-rate is 23B + D for North America and Japan, and 30B + D for Europe.

The discussion of AT&T plans for ISDN center on integrated access and on the signaling needed to support the resulting distributed processing. With this infrastructure in place, the next steps in the AT&T ISDN evolution can be identified. During the evolutionary phase, there will be compatibility at all levels, made possible by adherence to internationally accepted standards.

These steps in ISDN will accommodate the need for higher bandwidth and larger capacity. Higher bandwidths, of the order of 25 to 50 Mb/s, will arise from rapid and nearly universal high-speed fiber deployment. This higher bandwidth will accommodate high-quality video transmission and large-scale data transfer.

ISDN will allow access to today's muitiple networks on a call-by-call basis by using common access facilities. The integration of these functions to allow full flexibility requires a general-purpose network with a single fabric that can support all transport modes (circuit, channel, and packet). AT&T's wideband packet technology is capable of supporting virtual circuits, virtual channels, and individual packets. It also has a bandwidth capacity to handle the multiple service requirements of voice, data, and image.

The fundamental networking principles—networks based on distributed intelligence and open system interfaces—are critical for meeting future customer needs in advanced communications applications. Elements of this architecture are already in place in central office and premises-based networks. The upcoming ISDN offerings are the next logical step in the evolution of this architecture.

At least at first, voice seems to receive primary attention in any ISDN scheme. This is because transmission facilities are currently mostly analog, even though switching is increasingly digital. Despite its potential benefits, implementation of ISDN has been delayed by controversies among the various constituencies in the ISDN community, i.e., the carriers, regulators, investors, suppliers, and end users (Bolger 1988). For example, users have to be convinced that they really do need the enhanced services that can be provided by ISDN before they will make the corresponding financial commitment. Also, other approaches may be more appropriate in some applications. For example, technological criticism has raised the question of ISDN vs LANs (local area networks), PBXs (private branch exchanges), and FONs (fiber-optic networks). LANs, FONs, and PBXs (especially the first two), offer solutions that tend to favor data more than voice and may be more appropriate for local applications. They provide high bandwidth for data communications, which telephony network-based ISDN cannot presently do.

The first ISDN network became operational in West Germany in November 1988. It offers basic-rate service, with two 64 kbps channels for voice, data, and facsimile transmission (B-channels), along with a 16 kbps channel for control signaling (D-channel). Primary-rate service will provide 30 B-channels and one 64 kbps control D-channel. Numerous field trials are already in progress in the United States, Japan, France, and Great Britain.

Therefore, ISDN has made the transition from initial design to real-life implementation. However, although great effort has gone into the establishment of standards for ISDN operation, relatively little progress has been made in the development of analytical models to evaluate network performance. Although it is desirable to study optimality of integration for a "true" network of many nodes and links and of arbitrary topology, the difficulty (in fact, the enormity) of the problem has forced researchers to follow a moderate step-by-step approach. Thus analyses of voice/data integration systems have generally been confined to operation at a single node, or possibly at all nodes of simple network configurations. Simulation is required to evaluate all but the simplest scenarios. Section 7 reviews the movable-boundary method, which has been proposed for the integration of voice and data at a single node. Even for this simple case, very few results are available on control schemes that optimize network performance. Even fewer are available for multiple-node configurations.

ISDN in Military Environments

Like the commercial sector, the military has requirements for integrated communication services. The question naturally arises of whether the commercially developed ISDN is appropriate for military applications or if a completely different approach is necessary. We now summarize briefly some of the major issues associated with the use of ISDN in military applications. Our discussion is based in part on Silverman (1987). Other useful references include Roehr (1987) and Mercer and Edwards (1986).

Silverman (1987) concluded that most military communication requirements can be satisfied by the standard commercial ISDN. For example, the ISDN architecture will be able to provide high-quality secure voice because it uses 64 kb/s digital channels, provided that encryption devices are developed that are compatible with this signaling format. Another important issue is how well military networks will be able to exploit the advanced capabilities and protocol standards established for ISDN; this can be done, but it will not be a simple task. One important requirement not satisfied by the standard ISDN architecture is the need for a multilevel precedence and preemption (MLLP) capability. Silverman concluded that no problems are insurmountable in the transition of military communication networks to ISDN-based systems; however, considerable work is needed to develop a comprehensive security architecture.

Voice/Data Integration in Mobile Military Radio Networks

Section 4 outlines some of the major features that distinguish mobile military radio networks from the wireline or satellite-based networks for which integrated networking methods have been developed. The low data rates available in mobile military radio networks preclude the use of ISDN standards in these networks. However, the methods developed for ISDN, in particular the movable-boundary method for integrated switching, will serve as the starting point for our research.

6. SPEECH INTERPOLATION METHODS

A major goal of communication network design is to make efficient use of the communication resource. In particular, it is often desirable to maximize the traffic that can be supported by the network, subject to various constraints on performance, e.g., speech quality. Before we address the issue of integrating voice and data traffic, we consider ways in which voice conversations themselves can be multiplexed to improve channel efficiency.

The normal mode of operation for voice communication is a full-duplex circuit, i.e., two circuit-switched connections are established between the two parties in the conversation, one for communication in each direction. Clearly, this mode of operation does not make efficient use of the channel, because only one talker normally speaks at a time. Thus each channel is, on the average, idle at least half the time. Since there are also silent periods in conversation, speech is actually present on a telephone channel only about 40% of the time (Brady 1968); thus more than half of the channel resource is not used.

Several speech interpolation schemes (Campanella 1987) have been developed that permit a number of voice calls to share a smaller number of channels on a statistically multiplexed basis, thereby improving efficiency by a factor greater than two in some cases. The first such scheme was time-assigned speech interpolation (TASI), which was developed for analog channels. Improved performance can be obtained through the use of digital speech interpolation (DSI) methods; e.g., the digital implementation of TASI is often referred to as TASI-D. Other forms of speech interpolation exploit the inherent redundancy in speech, rather than simply the silence periods. Such schemes include speech predictive encoded communication (SPEC) and multirate voice processing. An overview of statistical multiplexing techniques for integrated voice and data traffic is given in Bially et al. (1980a).

This section discusses the principles of speech interpolation schemes and their potential applicability to integrated mobile radio networks in military environments. We note at the outset that our environment is quite different from that for which the TASI-type schemes have been developed. In particular, in our case only a small number of calls can be multiplexed onto the same waveform, whereas a large number are needed to obtain the impressive performance gains that can be achieved in commercial systems. Thus many of the results discussed in the literature cannot be applied directly to our problem, and future research is needed to develop appropriate speech interpolation schemes for the military radio environment.

Time-Assigned Speech Interpolation (TASI)

The TASI technique (Bullington and Fraser 1959; Fraser, Bullock, and Long 1962; Midema and Schachtman 1962) is used to permit the sharing of a number of channels (that share a wideband channel by means of FDM) by a larger number of users on a statistically multiplexed basis. The basic principle of TASI is quite straightforward. Speech activity detectors (SAD) are used to distinguish between talkspurts and silent periods. At the beginning of a talkspurt, the multiplexer assigns one of the currently unused channels to the speaker. This channel is relinquished at the end of the talkspurt. At the speaker's next talkspurt, an unused channel is again assigned. In general, the channels may be assigned at random from one talkspurt to another. The multiplexer sends the appropriate signaling information to the demultiplexer at the other end of the communication link to associate each talkspurt with the correct conversation.

Clearly, there is no guarantee that a channel will be available because of the bursty nature of the talkspurts. If a channel is not available at the beginning of a talkspurt, the initial part of the talkspurt is lost. This clipping of the initial portion of a talkspurt is called competitive clipping or freeze-out. Studies have shown that speech clips longer than 50 ms cause perceptible degradation of speech. Campanella (1987) discusses TASI under the criterion that the percentage of speech clips longer than 50 ms was less than 2%. The ratio of the number of input channels that can be supported to the number of channels available for transmission (subject to such a criterion) approaches the reciprocal of the channel activity as the number of channels increases; this ratio is somewhat greater than 2, provided that the number of channels is sufficiently large that the law of large numbers is applicable. A reasonable number of transmission channels for practical applications would be about 40 or more. When the number of channels is smaller, the interpolation advantage (also known as TASI advantage or channel multiplication ratio) that can be achieved is smaller.

Digital Speech Interpolation (DSI)

A digital form of TASI, sometimes referred to as TASI-D, can also be implemented. Upon its arrival at the multiplexer, each talkspurt is assigned a periodically recurring TDMA time slot (provided that one is available), which is analogous to the FDM frequency channel used in analog TASI.

TASI-D has several advantages over analog TASI. Digital speech detectors perform better than analog ones, and digital switching avoids the annoying clicks that characterize analog TASI. In addition, TASI-D permits the incorporation of variable rate and embedded coding techniques, which we discuss later in this section. Furthermore, voice in military communications is increasingly being encoded by digital rather than by analog waveforms because digital encryption for security reasons is both easier and less vulnerable to unauthorized decryption.

As with analog TASI, the number of channels availablc for transmission is a critical system parameter. In military communication systems, this number is quite low, since only low data rates (rarely greater than 9600 b/s) are typically supported. The data rate needed to support voice communication is somewhere between 2400 and 800 b/s, with possible reduction to 600 b/s, as we discuss later in this section. Thus the number of voice channels that can be multiplexed over a single communication link may be between 4 and 16, which is generally not sufficiently high for TASI operation.

During times of congestion, the number of bits used to specify amplitude can be reduced, thereby effectively increasing the number of calls that the channel can handle simultaneously. Alternatively, an embedded coding scheme can be used, under which the voice stream is divided into packets of two or more priority levels; lower priority packets would be dropped as needed (Kang and Fransen 1982; Bially et al. 1980a,b; Yin et al. 1987a,b). These approaches reduce the quality of the speech, but also reduce the occurrence of clipping because most of the high priority information will be delivered. Basically, these approaches spread minor degradation over all channels, rather than imposing severe degradation (clipping) over a few channels. In many applications, this is a reasonable trade-off. Performance of such schemes was evaluated by simulation only in Bially et al. (1980a) and by queueing analysis for a simplified model in Yin et al. (1987a,b). No effort to optimize in any sense was attempted. Note that the decision to drop packets would be made dynamically in response to instantaneous conditions at a node, thus resulting in a time-varying quality of speech. Also note that in multihop applications, the decision to drop low-priority packets could be made at intermediate nodes; of course, once the low-priority packets are dropped, they cannot be restored.

The inability to achieve high interpolation advantage when the number of channels is small may be alleviated by buffering the speech; this is possible in digital systems but not analog systems. Buffering by perhaps several hundred ms would reduce competitive clipping; alternatively, for a specified level of competitive clipping, the number of users able to share a fixed number of channels would be increased. Note that speech would have to be buffered not only at the originating node and relay nodes (i.e., when waiting for a time slot to become available) but also at the ultimate destination where buffering would be needed to maintain continuity over the duration of every talkspurt. Thus a trade-off exists between the TASI advantage that can be achieved and the delay that is tolerated. Basically, if buffering delays can be tolerated, a TASI advantage that is close to the reciprocal of channel activity can be achieved, even when the number of users is not sufficiently large for unbuffered TASI-D to provide good performance (Weinstein and Hofstetter 1979; Sriram et al. 1983; Janakiraman et al. 1984b). Buffering of speech by as much as several seconds may be a reasonable approach for military communications if this is the only way in which efficient channel operation can be achieved; further analysis of such systems is needed. However, communication resources can generally be used more efficiently if channel capacity is shared dynamically between voice and data needs by taking advantage of the fact that most data traffic does not have the same type of real-time delivery requirements as speech traffic.

We now address the impact of Navy waveform considerations on the performance of speech interpolation schemes.

Navy Waveform Considerations

We have noted that the number of signals that can be multiplexed onto a single waveform is a critical parameter in speech interpolation schemes. This number depends on the data rate supported by the channel and on the data rate required to support each voice call. The standard implementation of digital speech in the commercial sector uses 8-bit per sample pulse code modulation (PCM), resulting in a data rate of 64 kb/s. Data rates of 32 and 16 kb/s are also being considered for commercial applications. However, in the military environment, channel bandwidth is a scarce commodity, and the use of much lower data rates is necessary. We now review the characteristics of the encoding schemes that have been proposed for speech in navy systems, and we address their impact on system performance.

The linear predictive coder (LPC) technique is a practical approach for the implementation of digitized speech where large bandwidths are not available.* In particular, 2400 b/s LPC speech has been standardized within government agencies (Federal Standard 1015 or MIL-STD-188-113) and by NATO (STANAG 4198) (Kang and Fransen 1985). The 2400 b/s LPC is characterized by low quality, but highly intelligible speech (Kang and Fransen 1982). For example, it does not reproduce indistinct or rapidly spoken speech and is somewhat biased against female voices; for this reason Kang and Fransen (1985) have suggested that 4800 b/s LPC may be worth studying for some applications. Clearly, higher data rates result in better voice quality. However, since the overall data rates supported by military communication channels will be low (4800 b/s or possibly 9600 b/s), we are interested in using the lowest possible data rates so that some form of voice multiplexing or interpolation can be achieved. This may, in fact, be possible, because NRL has developed an 800 b/s LPC voice processor that provides performance that is only slightly worse than that of the 2400 b/s LPC processor (Kang and Fransen 1985). It is hoped that future research in speech processing will produce even lower data rate requirements for speech, perhaps as low as 600 b/s. Further reduction in data rate (to 300 b/s or lower) may be possible if conversation is limited to a predetermined vocabulary.

The ability to multiplex several voice streams over the same communication channel will make it possible to exploit some of the approaches that have been developed for voice multiplexing in commercial applications. However, the fact that only a small number of voice calls can be multiplexed over the same waveform will preclude the achievability of large interpolation advantages. We have noted that buffering speech (at the expense of increased delay and buffer storage requirements) may be a reasonable approach for some military communication applications. More research is needed to determine the performance of speech interpolation systems under such constraints. Similarly, more research is needed to determine the performance of voice/data integration schemes, such as the movable-boundary method to be discussed in Section 7, for small systems of this type.

Multirate Voice Processing and Navy Applications

We have noted that multirate coding can be used to permit a variable number of digital voice calls to share a communication channel. The NRL multirate processor (Kang and Fransen 1982) uses

*For example, LPC speech at 9.6 kb/s provided speech quality that is comparable to that of the continuously variable slope delta (CVSD) modulator operating at 16 kb/s, and LPC at 16 kb/s is comparable to CVSD at 32 kb/s (Kang and Fransen 1982).

the LPC principle to generate a voice waveform at three data rates simultaneously—2.4, 9.6, and 16 kb/s. The waveform actually transmitted would depend on the channel and transmission system characteristics, including possibly the level of congestion in the network. The 2.4 kb/s bit stream is a subset of the 9.6 kb/s bit stream, which in turn is a subset of the 16 kb/s bit stream. Thus lower rate data can be extracted from the higher rate data in applications where data rates have to be changed at intermediate nodes in multihop (tandem) networks. An important property is that end-to-end encryption can be maintained in this bit-stripping process. However, the practicality of this approach in mobile radio network applications is limited by the fact that the lowest data rate available may be the highest data rate that could be transmitted in the network. Thus operation at 2400 b/s would be commonplace, leaving no room for further lowering of the data rate.

7. VOICE/DATA INTEGRATION METHODS AT A SINGLE NODE

We have discussed the differences between voice and data communication, and the resulting differences in the networking schemes that must be developed to support them. For example, voice is characterized by a need for real-time delivery and by the need for a relatively long-term commitment of channel resources (typically several seconds to minutes). To support this type of traffic, either circuit switching or virtual circuit switching is generally used. This section discusses the problem of voice/data integration from the perspective of multiplexing voice and data at a single network node. The two possible control actions the multiplexer can take are to accept the voice call (thereby implying a long-term commitment of resources, i.e., until the call is completed) or to reject it.

On the other hand, data traffic is often bursty in nature, and messages may consist of one or a small number of packets. Low delay is usually desirable. However, since real-time delivery is not normally required, a store-and forward mode of operation with buffering is usually feasible. This results in a preference for the datagram mode of packet switching in many applications. Large data file transfers may also be considered, for which virtual circuit switching may be appropriate. However, unlike voice, delay is usually not an important performance criterion for such traffic, which may often be handled with a lower precedence rating.

The goal of voice/data integration is to share network resources efficiently between these two classes of traffic while satisfying the performance requirements of both. Networking schemes are needed that can simultaneously provide the equivalent of circuit switching for voice and packet switching for data. From the perspective of operation at a single node, a multiplexing scheme is needed that satisfies the real-time delivery requirements of a sufficiently large number of voice calls, while providing for adequate data throughput with acceptable delays. Our discussion focuses on the "movable-boundary" scheme, which has emerged as the prime hybrid system of integrated switching. The term slotted envelope network (SENET) is often applied to any boundary scheme for voice/data integration. However, in this report, we prefer to use the more general term movable-boundary to describe this type of multiplexing mechanism.

Boundary Schemes for Voice/Data Multiplexing

The most basic approach to voice/data integration is the "boundary" scheme for multiplexing, which is based on a TDMA frame structure* (Fig. 2). The fixed-length TDMA frame is partitioned into two compartments, with voice being circuit-switched in one and data being packet-switched in the other. The boundary between the two compartments can be either fixed or movable. In the fixed

*Alternatively, FDMA can be used to share the wideband channel.

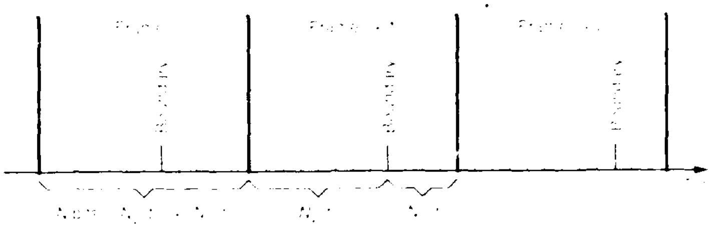

Fig. 2 — The movable-boundary channel

boundary scheme, voice is transmitted only in the slots of one compartment; any unused slots from the data compartment are left unused. Data are handled in a similar way. Under the movable-boundary scheme, data traffic is allowed to use any idle slots of the voice compartment, but not conversely. By permitting data to use the unneeded voice slots, higher bandwidth utilization is achieved.

The acceptance of a voice call by the multiplexer implies a long-term commitment of a channel (in this case a TDMA slot) to support the call. It is generally assumed that a voice call cannot be interrupted once it is assigned a channel. Data traffic requires only a short-term commitment, i.e., one packet at a time.* Since each data packet occupies only one slot at a time, data traffic does not interfere with voice traffic; the voice slot that was borrowed for data reverts to its original status as a voice slot immediately when needed for this purpose.

A great deal of flexibility can be achieved with movable-boundary schemes. For example, voice calls with different bandwidth requirements can be accommodated by assigning them slots of different lengths. Similarly, data packets of different lengths can be accommodated by assigning them different size slots. A control channel, which may be implemented as the first time slot in the frame, can keep track of changing multiplexer allocations. Other variations include the use of variable length frames and multiple boundaries (e.g., three classes of traffic may be considered, where the third could be long file transfers).

The use of movable-boundary schemes permits the use of dynamic optimization techniques that adapt to channel traffic. Based on traffic, the position of the boundary can be chosen to optimize system performance. Allocation of too small a fraction of the channel to voice traffic can result in the blockage of too many calls (voice traffic cannot use data slots, even when they are not in use); allocation of too large a fraction to voice can result in excessive queueing delay for data (or packet loss if buffer capacities are exceeded). In general, a performance measure can be defined as a weighted sum of voice and data criteria. Analysis of movable-boundary schemes is very difficult, except for the simplest of examples.

In the basic movable-boundary model, it is assumed that a slot is assigned to each call for its entire duration, even when the user is silent (i.e., between talkspurts). Section 6 notes that users are normally silent more than half the time, thus permitting the use of speech interpolation schemes that increase the number of users supported by a wideband channel. Digital speech interpolation schemes may be used in conjunction with the movable-boundary scheme to increase channel throughput. Alternatively, it is possible to transmit data in silent periods, a method known as time-assigned data interpolation (TADI). This section emphasizes the movable-boundary scheme but also indicates how these other methods can be used to complement it.

*Although data messages can consist of more than one packet, each packet of a multipacket message can be treated separately by the network since there is no need for uniformity of delay

Most performance evaluations to date have shown that movable-boundary schemes hold the greatest potential in integrated voice and data communication and that they have the greatest flexibility in handling various mix factors of voice and data. These mix factors range from one extreme of 100% voice (in which case it can become a pure circuit-switching scheme) to another extreme of 100% data (in which case it can coincide with a pure packet-switching scheme). Furthermore, movable-boundary schemes can be modified to accommodate buffering of voice and other dynamic allocation methods such as TASI and TADI. These schemes can be completely software-controlled; they can be easily reconfigured; and they are compatible with a local area network environment.

The main reason that the movable-boundary scheme possesses all these powerful advantages is that it is based on an inherently flexible and simple idea of implementing the switching function in the time domain by allocating an adjustable number of time slots to voice calls and data packets. Of course, it is not a panacea for all switching problems. It is, however, the best compromise for many network switching environments.

Movable-boundary schemes can be implemented by using either FDMA or TDMA structures. The analysis is easier for FDMA models because the resulting control problems are Markovian. TDMA models are more difficult to analyze because they yield semi-Markovian control problems. However, linear programming techniques, which have been applied to FDMA-based systems, may also prove to be useful for this type of problem (Viniotis 1988), as discussed in Appendixes A, B, and C.

Studies of the Movable-Boundary Method

The movable-boundary scheme has been studied extensively since it was first proposed by Kummerle (1974) and Zafiropoulo (1974). Since then, many variations of the movable-boundary scheme have been proposed and analyzed. We now review some of the major studies of these schemes.

Coviello and Vena (1975) described in detail the operation of a movable-boundary scheme developed for a T1 carrier, which they named the Slotted Envelope Network (SENET). The T1 carrier supports a data rate of 1.544 Mb/s; time is divided into 10-ms frames, each containing 15,440 bits. Each of these frames is divided into a large number of slots for voice and data applications. This scheme is very flexible in that a variety of data rates may be chosen for the voice and data slots, depending on the user demands. A common control channel (e.g., the first slot in each frame) is used to control the assignment and delivery of slots. The term SENET is often applied to any boundary scheme for voice/data integration. However, in this report, we use the more general term movable-boundary to describe this type of multiplexing mechanism.

Fischer and Harris (1976) were the first to analyze the movable-boundary scheme under the assumption of Poisson statistics for the voice and data inputs. They performed a detailed queueing analysis of voice and data traffic and presented results for channel utilization and delay. Weinstein et al. (1980) corrected an error in the analysis and demonstrated that the original analysis was too optimistic because large data delays can occur when the voice traffic load through the multiplexer exceeds its statistical average. This paper also addressed flow control mechanisms to reduce data packet delays, including control of voice bit rate, limitation of the data buffer size, and a combination of these methods. The effectiveness of these methods was demonstrated by simulations.

Konheim and Pickholtz (1984) generalized the assumption of Poisson statistics for the voice and data inputs to general renewal process statistics. Their model incorporates buffering of both voice and data. Lee and Un (1985) also studied several variations of the boundary scheme that permit the queueing of voice and data.

Since in a fixed-duration TDMA frame the capacity is not fuliy utilized when overall traffic is sparse, a variable-length frame model was considered (Miyahara and Hasegawa 1978; Maglaris and Schwartz 1981). In this scheme, the frame duration is adjusted in accordance with the traffic variations. To ensure synchronous operation of the circuit-switched voice calls, an upper bound on the frame duration is established and voice calls are buffered. By using heavy-traffic approximations, it was shown that the scheme indeed provides a savings in bandwidth.

Janakiraman et al. (1984a) also considered a variable frame-length multiplexer with a minimum-maximum constraint on frame length. By using a single-server queueing model, they showed that the performance of the multiplexer with respect to blocking, channel utilization, and delay is better than that offered by a fixed frame-length multiplexer.

Since many models do not lend themselves to exact analysis, a number of approximations based on fluid flow models of the data arrival and departure processes have been considered. Leon-Garcia et al. (1982) examined the movable-boundary scheme to estimate the mean number of data packets in an integrated communication system. They modeled the packet queueing process as a single-server queue with randomly varying service rate and infinite buffer space to obtain the quasi-stationary behavior of the process. When the conditions for obtaining the quasi-stationary behavior were not met, fluid approximations were used to obtain the behavior of the packet queueing process. Gaver and Lehoczky (1982) also used the fluid approximation to model movable-boundary schemes. They obtained the voice loss probability and mean data queue length as a function of the arriving traffic rate and determined how to allocate channel capacity between voice and data for fixed traffic intensities.

Ross and Mowafi (1982) evaluated the comparative performance of the basic variations of the movable-boundary scheme for multiplexing data and voice over a T1 carrier—the application for which the term SENET was originally introduced. They found that, in terms of voice delay, the hybrid SENET schemes showed less variation in voice delay with changes in throughput than did pure packet-switching schemes. Also, with respect to voice continuity, they showed that movable-boundary schemes consistently outperform packet switching schemes. They concluded that such hybrid schemes are preferred for integrated voice and data communication over purely packet-switching schemes and that, in particular, the movable-boundary scheme (either by itself or combined with dig.tal speech interpolation) appears to be the most promising scheme.

Arthurs and Stuck (1983) examined alternate schemes for allocating frame capacity to synchronous and asynchronous traffic by using a general queueing model. They studied the properties of the statistics of the amount of data waiting to be transmitted at the start of each frame.

Sriram et al. (1983) performed a discrete-time analysis of the movable-boundary scheme used in conjunction with digital speech interpolation (DSI), which is discussed in Section 6. DSI permits the number of active voice calls to be greater than the number of time slots assigned to voice in each frame (although some voice packets may be lost). The speech activity detectors used to implement DSI also permit the detection of unused slots that may be used for data traffic. Thus data may be transmitted in the voice slots that are 'eft over after the DSI operation is performed; the transmission of data in silent periods is sometimes called time-assigned data interpolation (TADI). The use of a discrete-time analysis is advantageous because it permits the modeling of a general distribution for the data message length (including fixed packet size), whereas the continuous-time analysis permitted the modeling of only exponentially distributed message lengths. Performance was evaluated under a variety of criteria, including probability of voice call loss, probability of speech clipping, probability of speech packet blocking, and data message delay.

In general, the literature on voice/data integration can be characterized by two features. First, virtually all attempts at analysis have been confined to isolated nodes of the network. Interaction from other nodes is too complicated and is ignored, except in some simulation-based studies. Second, almost all analyses have dealt with performance evaluation of ad hoc schemes. Very few studies have looked at integrated networking models from the perspective of optimal control. The first attempt to do so was by Maglaris and Schwartz (1982), who formulated the problem of dynamically allocating the bandwidth of a movable-boundary node to voice and data as a Markovian decision process (MDP). They showed that the movable-boundary scheme does, in fact, display near-optimum properties. Other efforts in this area may be found in Viniotis and Ephremides (1987, 1988a); Lambadaris et al. (1987); and Tsang and Ross (1988). In Viniotis (1988) and Viniotis and Ephremides (1988a) a first attempt to analyze a tandem network of movable-boundary nodes is made.

All these analyses are not conclusive. They are based on specialized traffic assumptions and on the use of inadequate analytical tools. Furthermore, the analysis has also suffered from several other limitations. First, it has been assumed that no buffering is required for voice (since classical circuit switching was used for the voice compartment), and that an infinite number of buffers was available for the data traffic. Second, buffer management schemes and their relationship to channel management schemes have been treated in only a cursory fashion. One finds many analyses of buffer sharing strategies or channel allocation strategies in isolation, at a single node. However, analyses of integrated buffer/channel sharing schemes are few.

Appendixes A, B, and C discuss optimization problems further. In particular, Appendix A formulates the voice/data integration problem at a single node. This formulation is generic and amenable to optimization. We show that it is useful to characterize the problem as an MDP problem, and we discuss three solution methodologies that have been used for such problems. These are:

- the derivation of optimality conditions from the dynamic programming equation (DPE) that is associated with the corresponding MDP,
- the use of sample path or stochastic dominance arguments, and
- the reformulation of the MDP as a linear program.

Appendix B demonstrates how almost any queueing control problem that can be formulated as an MDP can be converted into an equivalent linear program. Appendix C formulates the optimal control of a movable-boundary scheme as an MDP and demonstrates how linear programming can be used to obtain the structure of the optimai policy. The development of improved optimization models for single-node applications and their extension to multinode operation in radio networks will be an important part of our studies on voice/data integration.

We now review the models of hybrid integrated nodes that use the boundary method to multiplex circuit-switched voice and packet-switched data. We begin with a baseline description that addresses both the number of time slots allocated to each type of traffic and the number of buffers required. We then discuss specifically the fixed-boundary scheme, the movable-boundary scheme, and some variations.

Baseline Description of a Hybrid Integrated Node

To establish a baseline description of the switching problem, we introduce a simple generic model. This model will serve as a point of departure for considering the various alternatives for integration by using the boundary method. Figure 3 shows a single switching node that is assumed to be part of a larger distributed packet communication network. As shown, the switching node consists of three basic elements: a finite buffer (which is shared in some fashion by all arriving traffic), a nodal processor (which selects data from the buffer, performs a limited amount of processing on it, and forwards it to the next node), and an outgoing channel (which has a fixed transmission capacity and is assumed to be shared by all outgoing traffic).

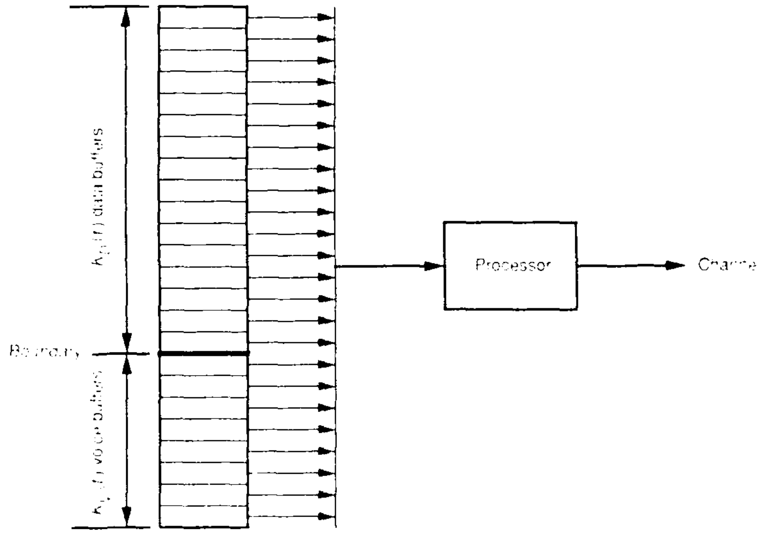

Fig. 3 — The movable-boundary node

Assume that the total number of buffers is K, the outgoing channel capacity is b bits per second, and the delay introduced by the nodal processor is 0. Furthermore, assume that at any given time t, the total number of buffers allocated to data traffic (denoted by $K_D\ (t)$) and the total number of buffers allocated to voice traffic (denoted by $K_V(t)$) are such that $K = K_D(t) + K_V(t)$. Data packets are lost when the buffer capacity is exceeded. We assume that the available channel capacity is divided into equal length frames of duration τ s each. Since the channel capacity is b bits/s, the frame size is $N = \tau\, b$ bits. Frames are further assumed to be subdivided into two compartments, one allocated to voice traffic and the other to data traffic. The size in bits, at any given time t, of these two compartments is denoted by $N_V(t)$ and $N_D(t)$, respectively, and is such that $N = N_D(t) + N_V(t)$. This is shown in Fig. 3.

The combination of the total number of buffers allocated to data traffic, the required nodal processing, and the total amount of channel capacity allocated to data traffic is called the data subsystem of the switching node. The voice subsystem of the switching node can be defined in an analogous fashion.

Voice and data traffic arrive at the switching node independently of one another. They are then either immediately buffered or lost; if buffered, they undergo some queueing delay before transmission over the outgoing channel. We assume that any queueing delay that may be experienced is caused by the limited channel transmission capacity and the effects of the channel allocation scheme. We neglect the delay introduced by the nodal processor.

Voice traffic is assumed to consist of a slowly varying, finite number of voice calls. Each voice call (VC) is assumed to be implemented as a packetized virtual circuit and is assumed to be independent of any other VC. Typically, the voice call interarrival and call length distributions associated with all VCs are assumed to be exponential, with parameters λ_{VC} and μ_{VC}, respectively. In such cases, since the voice call arrival process for each VC is Poisson, so is the sum of the arrival processes. Hence, the aggregate expected voice traffic arrival rate is given by $\lambda_V = n\lambda_{VC}$, and the interarrival distribution is exponential.

The switching node is capable of handling a maximum of n VCs. The selection of n is determined, in part, by the desired voice-call blocking probability and, in part, by the required minimum data-traffic throughput. When the maximum number of VCs has been attained, any new request to establish a VC is blocked until a currently active VC is deactivated; we assume that there is no queueing of voice calls.

Unlike voice traffic, data traffic is assumed to consist of individual packets; typically, these packets may arrive at the switching node according to an exponential interarrival distribution at an expected rate of λ_D packets per second. The expected length of all data packets is assumed to be $1/\mu_D$ bits; e.g., the packet length may be constant or exponentially distributed. Buffer space and channel capacity are assumed to be allocated to data packets on a first-come-first-serve basis.

The expected data and voice packet arrival rates are selected and enforced by the switching node's flow control mechanism. Selection of the expected data packet arrival rate λ_D for a given $K_D(t)$ and $N_D(t)$ is accomplished subject to the requirements that the expected loss rate of data packets from the switching node is less than a given loss-rate threshold.

Similarly, selection of the expected voice-packet arrival rate λ_V is assumed to be based on the current values of $K_V(t), N_V(t)$, and n, such that the expected voice packet loss rate stays below a chosen threshold and such that $\lambda_V \leq n\ \lambda_{VC}$.

Enforcement of the expected arrival rates is accomplished by blocking excess packets.

The flow-control mechanism of the switching node accomplishes selection and enforcement by using the integrated buffer/channel sharing scheme and external flow-control protocols. The integrated buffer/channel sharing scheme regulates the use and allocation of buffer and channel capacity within the switching node. External flow-control protocols inform other nodes of the successful or unsuccessful receipt of packets by the switching node and of the aggregate expected traffic rate that is currently available for each class of traffic.

Although voice traffic is typically given higher priority than data traffic for acquiring and retaining buffer space and channel capacity (up to the specified maximum number n of VCs), the objective of an integrated buffer/channel sharing scheme and network flow-control policy is to ensure that the rate of loss of data packets is maintained below the chosen threshold (or equivalently, that data-packet throughput is sustained above a specified threshold). As a consequence, the integrated buffer/channel sharing scheme is continuously presented with the conflicting goals of attempting to minimize VC setup time and voice packet losses while trying to maximize data packet throughput and minimize data packet loss, all within the constraints of finite buffer and channel capacity. The means by which each variation of the boundary scheme satisfies these requirements provides the key to measuring their relative adaptability and ultimate performance.

The Fixed-Boundary Scheme

The fixed boundary scheme divides the available channel capacity into equal length frames, each of duration α s. Since we assume that the channel capacity is b bits per second, the frame size is $n = b\alpha$ bits.

Each frame is further divided into two mutually exclusive compartments allocated to voice and data, respectively. The size N_V in bits of the voice compartment is *fixed* and selected so that voice call-blocking probability requirements are met. The size N_D in bits of the data compartment is also fixed and set equal to the remaining frame capacity; i.e., $N_D = N - N_V$. The mean voice packet length (for all VCs) is $1/\mu_V$ bits per packet, and the mean data packet length is $1/\mu_D$ bits per packet. Hence, the mean channel bandwidth allocated to voice traffic is $\mu_V N_V/\alpha$ voice packets per second, while the mean channel bandwidth allocated to data traffic is $\mu_D N_D/\alpha$ data packets per second.

The voice compartment is divided into a number of slots; each slot is of sufficient length to satisfy the bandwidth requirements of the particular voice call to which it is assigned. Given the mean channel bandwidth for voice traffic as previously described, there are $n = \mu_V N_V$ voice compartment slots. Since we assume a one-to-one correspondence between voice compartment slots and VCs, the maximum number of simultaneously active VCs that can be supported by the voice compartment is n. Obviously, we must have $\lambda_V \leq \mu_V N_V/\alpha$.

The static buffer-allocation scheme implies that the total number of buffers available within the switching node is partitioned into two fixed sets: one dedicated to voice traffic and the other dedicated to data traffic. Once made, the particular allocations are held fixed, and no sharing is permitted. By assuming that the voice bit flow is constant and because $\lambda_V \leq \mu_V N_V/\alpha$, the necessary and sufficient number of buffers that must be allocated to each VC to maintain uninterrupted continuity is *two*. Since a maximum of n VCs can be supported by the voice compartment of the channel, $2n$ buffers are required for voice traffic in the switch's buffer; i.e., $K_V = 2n$. Finally, if the total number of buffers at the switching node is K ($>2n$), the number of buffers allocated to data traffic is $K_D = K - 2n$.

It must be repeated that, under this scheme, the allocation of buffer and channel capacity between the voice and data subsystems, once made, is held fixed regardless of any variation in the number of active VCs, the expected voice packet arrival rate with each VC, or the expected data packet arrival rate. Hence, the voice and data subsystems can be viewed as operating independently of one another.

The Movable-Boundary Scheme

As with the fixed-boundary scheme, the movable-boundary scheme divides channel capacity into equal-length frames, each of which is further subdivided into two compartments: one for voice traffic and one for data traffic. Unlike the fixed-boundary scheme, however, although the buffer allocation remains fixed, the boundary between the voice and data compartments is not fixed. Instead, as the number of active VCs varies over time, the data subsystem may acquire and use those portions of the voice-channel capacity that would otherwise be unused as a result of fewer than the maximum possible number n of VCs being active. In particular, it is assumed that once a previously active VC is deactivated, its portion (i.e., slot) of the voice channel capacity is immediately allocated to the data subsystem for its use. It is further assumed that the switching node's flow-control mechanism immediately calculates a new value for $\lambda_D(t)$ and informs the network of this new value. The particular value for $\lambda_D(t)$ selected is governed again by the requirement that the loss probability remains below a given threshold.

On the other hand, when a new VC requests activation at the switching node, the channel capacity it requires is immediately reallocated from the data subsystem to the voice subsystem, provided that the new VC does not cause the total number of VCs to exceed the maximum n. As before, it is assumed that the switching node's flow-control mechanism immediately calculates a new value for $\lambda_D(t)$, which is actually achieved with a fixed constant delay.

The buffer-allocation mechanism for the movable-boundary scheme is identical to that for the fixed-boundary scheme. That is, if the total number of buffers at the switching node is K, then $K_V = 2n$ buffers are allocated to voice traffic, and $K_D = K - 2n$ buffers are allocated to data traffic. no buffer sharing is permitted, and the allocation, once made, is assumed to remain fixed.

As with the fixed-boundary model, the two queues are independent as far as arrivals and storage are concerned, but they are permitted to share service resources.

Exploitation of Silent Periods in Speech

In the basic movable-boundary model, it is assumed that a slot is assigned to each call for its entire duration, even when the user is silent (i.e., between talkspurts). Section 6 notes that users are normally silent more than half the time, resulting in inefficient use of the channel. The use of speech activity detectors permits the use of digital speech interpolation schemes that increase the number of users supported by a wideband channel. Digital speech interpolation schemes may be used in conjunction with the movable-boundary scheme to increase voice channel throughput by permitting more than n VCs to share the n slots in the voice compartment (possibly resulting in occasional lost packets).

As mentioned before, it is also possible to transmit data in silent periods, a method known as time assigned data interpolation (TADI). We have already noted that Sriram et al. (1983) analyzed the movable-boundary scheme in conjunction with DSI and TADI. The combined use of these methods results in the most efficient use of the channel.

The strategy of redirecting the blocked data traffic to the voice portion of the switching node is based on the assumption that both the buffer and channel capacity allocated to voice traffic are inherently underutilized. This is because speech typically consists of alternating sequences of talkspurts and silent periods. As a consequence, the voice subsystem of the switching node can be used to absorb temporarily any data traffic input to a level that can be accommodated by the new data packet channel service capacity.

As in the previous schemes, allocation of buffers between the voice and data subsystems, once made, stay fixed for all time. Unlike the previous schemes, however, data packets may share voice subsystem buffers during overflow periods.

Dynamic Buffer Allocation

The movable-boundary scheme can be considered with the additional capability of dynamic allocation of buffer space between voice and data within the switching node. Such a scheme would operate analogously to the movable-boundary scheme where the number of VCs is constant or decreasing. However, when the number of VCs is increasing, the required channel capacity is not reassigned immediately. Instead, it is delayed according to the following strategy.

When a virtual circuit is deactivated, not only is its associated voice channel capacity released for use by the data subsystem, but its associated buffers are also released. More specifically, if a single VC becomes inactive, its voice compartment slot and two buffers are assigned to the data subsystem. This buffer and channel capacity assigned is assumed to occur immediately upon deactivation of the VC.

On the other hand, when a new VC requests activation, this scheme immediately removes the equivalent of one voice slot and two buffers from further use by the data subsystem. However, before reassigning these resources to the voice subsystem, any data packets that might reside in the buffers must first be served. Once these data packets have been served, the buffer and channel capacity reassignments are made and the virtual circuit is activated.

Note that the practical reason for introducing this VC delay is that a non-zero probability exists that data packets will have accumulated in the buffers that have been designated for reassignment to the voice subsystem; this delay ensures that data packets, once accepted by the switching node, are not rejected or lost.

Mixed SENET

The mixed SENET scheme (Upton, 1984) is based on the movable-boundary scheme for channel allocation and dynamic partitioning for buffer sharing with several additional features.

First, when a VC is deactivated, its associated voice channel slot is immediately released for use by the data subsystem. However, the VC's associated buffer space is released only if doing so would not lower the number of reserve buffers below a prespecified number. If m is the number of currently active VCs, then $2m$ buffers are required to support them. Let η be an even integer that specifies the number of buffers the voice subsystem tries to maintain in reserve; i.e., buffers that are not currently dedicated to an active VC. Then, a deactivated VC's associated buffer space is released for use by the data subsystem only if, upon deactivation of the VC, the total number of buffers contained within the voice subsystem is greater than $\eta + 2m$. Note that, as a result of this condition, if we assume there are initially $2n$ (the maximum possible) buffers supporting n active VCs within the voice subsystem of the switch, then only after η VCs become inactive may voice subsystem buffers begin to be released to the data subsystem. Furthermore, at least 2η buffers will always be in the voice subsystem regardless of the number of active VCs.

A second feature of this scheme is that when a new VC requests activation, the channel capacity it requires is immediately reacquired from the data subsystem. Like the previous scheme, this scheme immediately removes the equivalent of one voice channel slot and two buffers from further use by the data subsystem. However, before reassigning these resources to the voice subsystem, any data packets that might reside in the buffers are first served. Once the data packets have been served, the buffer and channel capacity reassignments are made, and the virtual circuit is activated.

Finally, if the number of reserve voice packet buffers falls below η for any reason, the switch's flow-control mechanism readjusts the data traffic arrival rate. This reassigns a sufficient number of data subsystem buffers to the voice subsystem to bring the number of reserve buffers back to η. This is accomplished in a fashion identical to the way in which delayed VCs are handled.

Final Remarks

We have noted that research on voice/data integration has concentrated on the multiplexing operation at a single network node because of the difficulty of evaluating multinode network configurations. The movable-boundary scheme (combined with some form of digital speech or data interpolation) has emerged as the method of choice because it supports both (virtual-) circuit-switched traffic and packet-switched traffic, and it permits the incorporation of dynamic control schemes. Tables 1 and 2 compare the hybrid boundary-type buffer/channel sharing schemes we have described in terms of their key parameters and features.

In this study we plan to develop improved analytical models that permit optimization, both at a single node and on a networkwide basis. These models must reflect the nature of the mobile radio network channel, as described in Section 4. Section 9 addresses some of the issues associated with the control of multinode networks.

We conclude this section by noting that there are other ways to integrate data with voice. Appendix D discusses a voice/data integration scheme that is based on the 2400 b/s LPC system. An approach such as this one may be useful in augmenting the communication capability that may be achieved between a single source-destination pair, and it can be used in conjunction with the hybrid switching techniques described in this section.

Table 1 — Key Parameters of the Boundary Schemes

Scheme	Parameters			
	Voice Subsystem Arrival Rate	Data Subsystem Arrival Rate	$N_{V,D}$	$K_{V,D}$
Fixed boundary	$\lambda_V(t)$	$\lambda_D(t)$	Constant	Constant
Movable boundary	$\lambda_V(t)$	$\lambda_D(t)$	Time dependent	Constant
Movable boundary with DSI	$> \lambda_V(t)$	$\lambda_D(t)$	Time dependent	Constant
Movable boundary with dynamic buffer allocation	$\lambda_V(t)$	$\lambda_D(t)$	Time dependent	Time dependent
Mixed SENET	$> \lambda_V(t)$	$\lambda_D(t)$	Time dependent	Time dependent

Table 2 — Key Features of the Boundary Schemes

Scheme	Features				
	Buffer Allocation	Buffer Sharing	Channel Allocation	Channel Sharing	VC Set-up Delay
Fixed boundary	Static	No	Static	No	No
Movable boundary	Static	No	Dynamic	No	No
Movable boundary with DSI	Static	Yes	Dynamic	Yes	No
Movable boundary with dynamic buffer allocation	Dynamic	No	Dynamic	No	Yes
Mixed SENET	Dynamic	Yes	Dynamic	Yes	Sometimes

8. CONTROL OF INTEGRATED VOICE/DATA NETWORKS

Section 7 discusses the problem of voice/data integration from the perspective of multiplexing at a single node and notes that few analytical results on optimal control have been obtained. When the problem of networkwide control is addressed, the problems become considerably more difficult. This section describes some of the most notable and representative approaches that have been proposed in addressing this problem. Although many attempts have been made in the last 10 years, very few substantive results have been obtained to date on networkwide control of integrated voice/data networks. Perhaps this disparity between interest and achievement indicates that the "right" approach to the problem has not been discovered yet. A useful approach that holds some promise is one that looks at tandems of nodes that correspond to virtual circuits as the building blocks of a general network. Results on scheduling for real-time communications that have been obtained recently by Bhattacharya and Ephremides (1988, 1989) provide some motivation for this argument, which finds additional support emerging from the vast literature on voice/data integration. Section 10 discusses the problem of communication under time constraints.

The problems of routing and flow control in an integrated network cannot be decoupled from the basic multiplexing problem at a single node, which was discussed in the preceding section. First, the multiplexing function itself (whether or not implemented in terms of the movable-boundary method) is a form of flow control, since it performs a first screening function on new incoming calls. Second, the routing function, by virtue of selecting a virtual-circuit path for each voice call, implies the commitment of bandwidth (hence, affecting the multiplexing function) in subsequent nodes and links of the network. This intricate coupling of the controls at different locations in the network is the primary source of complexity that has retarded the advancement of the field. Thus, not unexpectedly, most of the work on the subject has been ad hoc and experimental, rather than analytical.

Network Control Schemes

Before discussing schemes for integrated networks, we first consider a scheme developed strictly for speech transport in mobile packet radio networks. Shacham et al. (1983) considered a single-channel network, i.e., a network that supports only one packet stream at a time in any neighborhood; thus the problem of multiplexing is not relevant here. A "duct-routing" protocol was developed under which a primary route (normally traversing several intermediate nodes) is established between a source and destination pair. When communication on one of the links on the primary route is unsuccessful (e.g., as a result of topological change, equipment malfunction, or jamming) a "helper" node routes the packet along an alternate path. Control mechanisms are used to avoid flooding the network with too many copies of the same packet. Experimental results show that the duct-routing protocol provides better speech quality than the regular data protocol that had been used for the network.

Voice/data integration may be considered for low data-rate applications in which the voice signal uses the complete bandwidth of the transmitted waveform, thus precluding the multiplexing of additional voice or data signals. In such cases, voice traffic could be handled only if the data transport operation is disabled throughout the duration of each voice call. Although the military waveforms being considered in this study can support only low data rates, they should be able to support some degree of multiplexing, as discussed in Section 6. Thus we are primarily interested in applications where multiplexing of voice and data is in fact possible, although limited to relatively small numbers of signals.

As noted earlier, the use of virtual circuits is usually the preferred switching method for voice because of the need to provide real time delivery of speech. The establishment of virtual circuits implies a commitment of resources (time slots or bandwidth units) by all nodes along the path, but only in a statistically multiplexed sense. Thus, as a result of normal traffic fluctuations, some packets may be lost. To alleviate this problem, embedded coding schemes may be used, under which the voice stream is divided into packets of two or more priority levels, as discussed in Section 6. Lower priority packets can be dropped during periods of high congestion with little degradation in speech quality, resulting in an implicit form of flow or congestion control. Performance of such schemes for single-node operation was evaluated by simulation only in Bially et al. (1980a) and by queueing analysis for a simplified model in Yin et al. (1987a,b). No effort to optimize in any sense was attempted. It would be desirable to be able to model such schemes in multihop applications. However, as noted in Section 6, the applicability of such schemes to military mobile radio networks may be limited. This is because in normal operation integrated networks may already be using the lowest possible data rate that can support acceptable speech quality, so that some degree of speech multiplexing and/or voice/data integration is possible. Thus there may be no room to back off to lower data rates for voice.

In Li and Mark (1984) the idea of window flow control was borrowed from nonintegrated environments and was explored for time-division multiplexed integrated systems. Again, the effort consisted mostly of microanalysis based on gross averages and approximations, and did not (or perhaps, could not) attempt optimization.

In Douligeris and Mazumdar (1988) the performance index of network "power" (defined as the ratio of throughput to average delay) was chosen and studied in the context of integrated environments. This study was based on the assumption that each commodity (i.e., the voice and the data traffic processes) desires to optimize its own network power. Thus a game-theoretic formulation was proposed, and an interesting analysis was pursued that searched for Pareto-optimal and Nash equilibrium solutions. Although this approach deserves credit for its ambitious multioptimization objectives, it fell short of its goals. In part hampered by analytical difficulties, but mostly owing to poor modeling, the approach did not match the problem properly. It failed to account for the special characteristics of voice traffic, namely the need for low and invariant delay, and the ability to withstand errors and losses (a shortcoming that is shared with many "multicommodity" approaches to integration).

In Ibe (1985) and Yum and Schwartz (1987) more pragmatic approaches were taken. Ibe formulated the routing problem as a multicommodity flow problem based on macroscopic, steady-state-type flow models (in the spirit of the original Gallager algorithm (Gallager 1977)). The cost functions were chosen to be the mean and the variance of the voice-packet delay. A static optimization procedure was then followed, based on a standard, Lagrange multiplier method. Yum and Schwartz studied the effects on the circuit-switched (data) average delay of the trade-offs involved in the selection of routing rules for purely circuit-switched, nonhierarchical networks (something that had been previously studied by Yum and Schwartz as a separate problem), where the movable-boundary idea was used as a means of switching integration.

A (philosophically) similar approach was taken in Maglaris et al. (1987), where a realistic integrated model was considered in which digital speech interpolation was used for voice switching and low-precedence, cut-through switching was used for data. Again, the analysis was based on average flows and simplifications. However, several practical issues such as stability and loop freedom were addressed as fallout from the resultant linear program solution.

In Lippman (1985) a simulation study of three poorly described alternative routing algorithms was performed for an integrated model of the Defense Switched Network. Thomopoulos and Zhang (1987) partially analyzed three alternative flow-control methods (which were inappropriately compared with the movable-boundary method, which is not exactly a flow-control procedure, nor is it easy to analyze across more than a single node in a network).

Perhaps the most notable (and hopeful) advances in the area of integrated, networkwide flow control and routing are the ones described in Ross and Tsang (1989) and Katevenis (1987). Ross and Tsang considered a simple flow-control model in which messages (or calls) are accepted or rejected at a single node (but for circuit-switched operation, which implies networkwide operation). A fixed number of classes of traffic is considered without attempt to distinguish the special characteristics of voice and data (or other real-life commodities, such as video, or high-priority tactical data). However, the approach is analytically rigorous and poses optimization goals. The solution is carried out by means of Markovian decision problem formulation tools (such as the ones described in Appendix A), and threshold-type results are obtained by direct use of dynamic programming methodology.

Katevenis, on the other hand, proposed a radically different and potentially far-reaching approach to flow control. This approach is based on the availability of new technologies and is especially applicable to wideband networks (such as SHF or EHF satellite systems and/or fiber-optic media), although not to HF networks or other applications in which the channel resource is a scarce commodity. It assumes superfast switches that are based on parallel (VLSI-implementable) architectures and plentiful availability of inexpensive, compact memory. Then a rotary switching concept is introduced that samples all virtual circuits at their required rates, thereby eliminating the queueing delay. The idea is simple (yet brilliant) and might represent a solution concept for the future, provided the assumed availability of technology matches the requirements of bandwidth, response time, etc.

The movable-boundary method for a tandem of nodes was recently studied by Viniotis and Ephremides (1988a), who used a Markovian decision process model to formulate the problem as a linear program. They showed that the optimum policy, which minimizes the weighted sum of blocking probability and data delay, is characterized by a switching curve. Appendix C discusses this approach in greater detail.

Final Remarks

The approaches we have outlined in this brief survey represent only a small (yet carefully selected to be significant and somewhat representative) portion of the work effort that has gone into the question of integrated network design. If anything resembling a trend emerges from this work, it is perhaps that the network must not be thought of as consisting of nodes and links that can be arbitrarily interconnected and switched. Rather, it should be thought of as consisting of arbitrary virtual circuits (that, of course, do share links and nodes among them), each of which, in turn, can be thought of as a tandem of links. Then, the transition from single-node analysis, as described in the preceding section, to full-network design, starts to appear feasible (though, still, only remotely so). To demonstrate somewhat more concretely that this is indeed a step in the right direction, in Section 10 we outline briefly an approach that concerns real-time communication (and is thus applicable to packetized voice) and addresses the problem of scheduling the transmission of messages that are subject to strict deadlines. Before we do so, however, we address the problem of channel access in integrated radio networks.

9. CHANNEL ACCESS IN INTEGRATED RADIO NETWORKS

This section discusses the major issues associated with channel access in integrated networks and presents our preliminary thoughts on this subject. Section 4 notes that the channel-access mechanism is a critical aspect of the design of radio networks. The process of channel access induces a much greater degree of interaction among network nodes than is present in wireline networks, and thus creates a totally different networking environment. In general, channel-access protocols must reflect the characteristics of the communication traffic that must be supported as well as the nature of the communication medium. In particular, when considering integrated voice/data radio networks, channel access methods must be developed that reflect the different requirements on delay and error rate that are associated with voice and data traffic as well as the impact each type of traffic has on the other.

The traffic generated at any node may consist of either a single user's transmission or a multiplexed waveform containing the traffic of several users. Also, the traffic may consist of either voice only, data only, or a combination of the two. In this section we do not distinguish separately each of these possibilities, but we limit our discussion primarily to general issues of channel access for voice and data. These special cases are topics for future research. At the end of this section we discuss the importance of studying the interrelationships between channel access and other aspects of network design.

Channel Access for Voice

Voice traffic, which is tolerant of relatively high error rates, is characterized by the need for delivery as a continuous stream in near-real time. This discussion assumes that buffering is not possible. Calls are blocked if channel resource is not immediately available, and the transmission of a voice call requires a continuous commitment of channel resource (e.g., a time slot in every frame) for the entire duration of the call. Data traffic may be either bursty or regular in nature and may consist of either one or several packets. In either case, the data are characterized by a need for very low packet-error probability but not for ·eal-time delivery. Delay requirements depend on the nature of the traffic and may be different for different classes of traffic in the network. Buffering of data packets is permitted.

The need to support the long-term nature of voice traffic results in a requirement for contention-free channel access once a call has been set up. However, traffic requirements change as existing calls are completed and as new calls are established. Despite the fact that voice calls are generally of long duration, as compared with data packet size, voice traffic is normally quite bursty when viewed from the perspective of call arrival and completion times. A Poisson arrival process is often used to describe call arrivals, and an exponential distribution is used to describe call duration. Thus fixed schedules cannot normally be used for voice traffic, and the protocol must be able to adapt to changing communication requirements by means of a demand-assignment scheme.*

Reservation schemes, which can maintain throughput levels near channel capacity, are the logical choice for voice calls. Many reservation schemes are proposed in the literature, and most of them are modifications of the demand-assignment scheme originally proposed by Roberts (1973). Reservations may be made either on a contention basis or contention free; in the latter case, a reasonable approach is to use the schedules produced by a link-activation algorithm (such as those developed by Baker et al. (1982, 1984)) as a reservation channel. Once a reservation is successfully made and acknowledged, the user is allowed contention-free access to the channel until the end of the call. This is signaled by the user's end-of-message indicator, at which point the channel (time slot) becomes available for assignment to another user. The reservation mechanism can be implemented in either a centralized or decentralized manner. However, decentralized reservation schemes are highly vulnerable to inconsistencies in the databases at the users, potentially resulting in chaos if two users think they are supposed to transmit at the same time. Such inconsistencies can be expected to be commonplace in hostile military environments. Note that, unlike most reservation schemes for data traffic, there is no need to specify the call length a priori. This is because any time lost in turning the channel over to the next user is negligible in comparison to the length of the call. In contrast, in data applications where messages usually consist of a small number of packets, the message duration is normally specified to minimize such wasted time and thereby maintain high throughput. The impact of propagation delay is especially important in geosynchronous satellite networks, which are characterized by a round-trip propagation time of about 0.25 s.

In multihop applications it is necessary to reserve a channel for each call on every link of the multihop path before transmission of the call can actually begin. Thus the channel-access problem is intimately related to the routing problem, since routing decisions directly affect the amount of congestion (and hence availability of time slots) throughout the network. Before addressing the use of integrated channel-access protocols, however, we make a few more observations on the voice channel-access problem.

The problem of voice-call channel access bears some similarity to the problem of multipacket-message channel access. For example, Reservation-ALOHA (Lam 1980) is a hyb 'd scheme under which only the first packet of the message contends for channel access. Unaer this scheme, once the

*Fixed schedules may be appropriate in some applications such as those in which a large number of users are attached to each node (thus the total demand summed over all users sharing the node may be relatively constant) or in those in which the communication resource is sufficiently ample to support many high capacity links so that efficiency is not of critical importance.

first packet is successful the user is granted ownership of one time slot per frame as long as there are packets to transmit, i.e., for the duration of the call. Such an approach might be considered for single-hop voice call applications. However, in multihop applications a time slot is needed for each hop along the way. This approach may result in the initial acceptance of a call for transmission over the first link but the subsequent abortion of the call if a collision is experienced downstream. Another disadvantage, even in single-hop applications, is that the call might have to be started several times before it successfully acquires a time slot. Use of a conventional reservation scheme appears to be more appropriate, especially if contention-free reservations can be made. Even when only contention-based access to the reservation channel is possible, the use of smaller reservation packets will permit efficient use of the reservation orderwire channel.

Channel Access for Integrated Voice and Data

Many approaches are available for channel access in data networks. Descriptions of many of these schemes may be found in survey articles by Tobagi (1980) and Lam (1983). Depending on the nature of the traffic, either contention-based or contention-free schemes or their hybrids can be used. Many comparative discussions of channel access protocols are in the literature, and we do not intend to discuss their merits here. However, the problem of channel access in integrated networks has not received much attention in the literature, and it is worth mentioning the models studied by Suda et al. (1983) and Wu and Li (1988) for access to satellite channels. These schemes are characterized by a fixed-length frame structure that contains reservation channels and information channels (some of which are allocated for voice and the remainder for data). Voice is handled on a reservation basis in both of these studies. Once a reservation for a voice call is made successfully, one slot per frame is allocated to the call until its completion. There are a number of differences in the two models, however, particularly in the channel access mechanism for data.

Suda et al. considered only the use of the slotted ALOHA random-access protocol for data traffic (making the assumption of infinite buffer capacity) and considered operation only under a fixed-boundary scheme. They noted that extensions of their analysis to movable-boundary schemes and finite buffer models would be difficult. Wu and Li considered three options for data traffic, namely, random access, reservation, and hybrid (i.e., combining features of random access and reservation).* Where reservations are used (i.e., for all voice schemes and for random access and hybrid data schemes) a distributed reservation scheme was used. Both fixed- and movable-boundary schemes were considered for sharing the channel resource between voice and data traffic. However, their analysis was overly simplified.

Clearly, similarities exist between the channel access problem and the multiplexing problem that are discussed in Section 7. However, there are also some important differences. In the multiplexing problem the goal is the optimal sharing of a contention-free channel that has already been established between the source and destination. A single node makes the decisions of which voice and data streams to multiplex into a single waveform. The link throughput can be maintained at its maximum value as long as the node has something to transmit over it. An important trade-off in mutiplexing is that of call-blocking probability vs data-packet delay (data-packet loss can also be considered in the case of finite buffers).

*The distributed reservation scheme considered by Wu and Li is similar in principle to the Interleaved Frame Flush Out (IFFO) protocols developed by Wieselthier and Ephremides (1980) for data. The class of IFFO protocols contains both pure reservation and hybrid (random access and reservation) schemes.

In the channel-access problem, the goal is the optimal use of a channel by a distributed population of users that can interfere with each other's transmissions. The channel-access problem is considerably more complicated than the multiplexing problem because not all transmissions are successful and because distributed operation is often necessary to obtain robust and efficient performance.

Spread-Spectrum Considerations

The use of spread-spectrum signaling, necessitated by antijam considerations in many military applications, leads naturally to the use of code-division multiple-access (CDMA) techniques. The use of CDMA provides an environment that is radically different from that of narrowband time-domain operation. A great deal of flexibility in channel-access protocol design can be achieved by taking advantage of the multiple-user capability, the selective-addressing capability, and the selective-reception capability of CDMA signaling (Wieselthier 1988). In particular, our future studies will exploit the ability of the CDMA channel to support several transmissions (on quasiorthogonal frequency-hopping patterns, or codes) simultaneously. The number of users permitted to share the wideband channel simultaneously can be described by a threshold corresponding to the allowable packet-error rate.* Voice traffic may be able to tolerate higher levels of other-user interference than data traffic because of its ability to tolerate relatively high error rates (as a result of its inherent redundancy). Voice traffic would have to take precedence over data traffic (in the sense of not permitting data to preempt ongoing voice calls) for the reasons discussed earlier. If the allowable threshold is not exceeded by voice calls alone, data packets may be permitted access to the channel. The decision of whether or not to permit a new voice call to access the channel may, in general, be based on a criterion that is a weighted sum of voice and data error probabilities that reflects the need to handle both forms of traffic. Note, however, that since data packets have more stringent packet-error probability requirements than voice, the threshold for data will be lower than that for voice (if the same code rates are used for both).

Relationships Between Channel Access and Other Network Functions

The relationships between routing and channel access must be investigated. It is easiest to address this problem when contention-free channel access is used. Clearly, these network functions are not independent, because the schedules used to govern channel access determine when particular links are active and thus determine the amount of traffic that can be transmitted through these links.† The link-scheduling function determines the capacities of the links in the network, one of the most critical factors in the generation of routing schemes. Conversely, the routing scheme impacts directly on the traffic that must be supported over each of the network's links. A joint routing-scheduling problem can be formulated as a two-stage optimization problem. First, for a given routing scheme, the schedule is determined that provides optimum performance. Then this performance is optimized over all possible routing schemes. Some preliminary results are available for the joint routing-scheduling problem for data-only applications (Tassiulas and Ephremides 1989). Extending these results to integrated radio networks would be of great interest. This problem is also of great interest in the study of mixed-media networks, i.e., networks that contain links of markedly different characteristics such as data rate, error rate, delay, or connectivity. Such studies represent one of the first

*The modeling of the effects of other user interference in CDMA systems is actually quite complex. Binary threshold channel models are often used. In these models, a packet is assumed to be received correctly as long as the number of other users transmitting simultaneously does not exceed the threshold; it is assume to be received incorrectly whenever the threshold is exceeded. Such models do not characterize system performance accurately (Wieselthier 1988). However, a legitimate use of a threshold based model is to characterize a link in terms of whether or not the specified packet-error probability is exceeded.

†The number of times a link is activated per frame determines the capacity of that link.

attempts to bridge the layers in the Open Systems Interconnection (OSI) seven-layer protocol model, i.e., to combine the channel-access function (normally considered to be part of layer 2, the data link layer) with the routing function (layer 3, the network layer). The OSI model is described in detail by Tanenbaum (1988). The design of a layered protocol for integrated voice/data networks is discussed in Hoberecht (1983).

Also, the impact of mutiplexing on channel access and routing must be studied. The quantity and type (voice vs data) of information multiplexed at a node will impact the channel-access performance that must be supported. Clearly, routing decisions affect the quantity and mixture of traffic at each node, thus complicating this problem further.

Future Channel Access Studies

This section has outlined the major issues associated with channel access in integrated radio networks. Despite the great interest in channel access over the past 15 years, and despite the great interest in voice/data integration, in fact, few studies have been made of channel access in integrated radio networks. This remains an area requiring fundamental research.

Channel access will be an important aspect of our voice/data integration studies. As discussed in this section, reservation schemes will be our primary choice for voice traffic. A variety of schemes, including scheduled access, demand access, and random access will be considered for data traffic. The use of fixed- and movable-boundary schemes for channel access will be investigated. The impact of military communication considerations will be addressed, including the use of spread-spectrum signaling.

We have seen that the problem of channel access should not be studied independently of routing because of the strong relationships between these two aspects of network control. Therefore, we will study jointly the channel access and routing problems, an area that is especially important for voice traffic because of the need to support virtual circuits throughout the duration of calls. We will also study the impact of multiplexing on both channel access and routing.

10. COMMUNICATION UNDER TIME CONSTRAINTS

As noted throughout this report, speech communication is characterized by the need for real-time delivery, which translates to a requirement for low delay and low variability of delay. Also, some data communication is time-critical in nature. This section addresses communication under time constraints, a crucial aspect of both voice and data communication.

First, we consider contention-free operation over a single network link, subject to a constraint on the time of message delivery. Such a system can be modeled as a single-server queueing system that accepts packets from neighboring links for transmission, each carrying its own "extinction" time. If service (transmission) of a packet does not begin by the extinction time of that packet, it is dropped from the system, since its delivery offers zero reward while occupying network resources that could be used for some other purpose. This is a reasonable model for voice communication, which is characterized by the loss of packets that are not available when needed. We are interested in minimizing the number of lost packets.

The problem becomes considerably more complicated when multihop network operation is considered. In this case, the time by which a packet must be received over a multihop virtual circuit path is specified. Factors that must be considered include the variable queueing delay that may be experienced at intermediate nodes and the interaction among the different traffic streams supported by the network, which are controlled by the various protocols that govern network operation.

Few results have been reported in the literature on problems involving deadlines. In fact, most of the work in the past has considered slightly different variations of the problem discussed above. This problem has been formulated either as a waiting-time analysis of systems with impatient customers (no scheduling) on a first-come-first-served discipline, or as the scheduling of jobs in closed systems (no arrivals or losses). The performance criterion of average tardiness (i.e., the time past the extinction time until the completion of the service) is considered, rather than the percentage of completed services.

We now summarize some results on optimal scheduling with strict deadlines that have recently been obtained by Bhattacharya and Ephremides (1988, 1989) by using stochastic dominance arguments (see Appendix A). We first consider a single link. Each message, upon its arrival, announces its extinction time. If transmission does not begin by this time, the message is considered lost and is never scheduled for service. The objective is to schedule the transmission of messages so that the average number of messages lost over any time interval is minimized. Under a fairly general set of conditions, including exponential service times (message lengths), it has been shown that the policy of scheduling the eligible customer with the shortest time to extinction (STE) is optimal among all nonpreemptive and nonidling policies.* This approach of scheduling the most urgent message first is certainly intuitively appealing. Sometimes (e.g., loosely speaking, when the only packets in queue have very long extinction times) better performance can be obtained when idling is permitted (i.e., by remaining idle in case a packet with a short extinction time arrives). In this case, the optimal policy among the class of nonpreemptive policies is in the class of STE with idling (STEI) policies; whenever the server is not idle, it schedules messages according to the STE rule. Bhattacharya and Ephremides also addressed the problem for the case in which the constraint is on the complete transmission time rather than on waiting time, as well as the case in which the scheduler does not have exact information on the deadlines. It would be desirable to be able to extend these results to other than exponential message lengths.

It is difficult to extend the analysis to the case of a tandem of nodes. To permit the analysis, Bhattacharya and Ephremides had to remove the condition that late messages are lost. Instead, a tardiness cost was considered, representing the delay beyond the extinction time with which the message is received. Under fairly general conditions it was shown that the policy of scheduling the message with the earliest deadline at each node minimizes the tardiness cost over a finite operating horizon among the class of nonidling nonpreemptive policies.

These studies show that communication under time constraints represents a challenging analytical study. For single links, further study is needed to extend these results to nonexponential service times. For tandem operation, it would be desirable to obtain results for the performance measure we are really interested in, i.e., the probability of packet loss. Ultimately, we would like to be able to consider more general network configurations. The study of these configurations raises a host of new problems such as the following.

*By eligible, we mean a customer whose extinction time has not yet occurred. Nonidling means that the server must never remain idle when there is an eligible packet awaiting service.

In radio networks the relationships between the multiple-access protocol and the scheduling policy (which controls the order in which packets are transmitted by a particular node) must be examined when dealing with time-critical applications. In particular, in a contention-based random-access environment, the service time (the time until the packet is successfully received at the destination) can be highly variable because collisions often necessitate packet retransmission. It may be necessary to maintain throughput at low levels to ensure that a sufficiently high percentage of packets are received within the specified time. Even in contention-free applications, packet errors may occur as a result of noise, jamming, or (in the case of CDMA applications) other-user secondary interference. Also, since the service time is variable, it may sometimes be best to throw away packets that have only a small chance of being received in time, with the goal of maximizing the percentage of packets that are, in fact, received in a timely fashion. Only a few studies of multiple-access protocols under time constraints have been made. These include Kurose et al. (1984, 1988); Panwar et al. (1987); and Paterakis et al. (1987). Of course, this problem becomes considerably more complicated when multihop operation must be considered.

Section 9 notes that the relationships between routing and channel access must be investigated because these two aspects of network control are intimately related to each other. There is no reason to believe that optimum performance can be achieved by addressing these problems separately. In this section we are again primarily (although not exclusively) interested in contention-free link-activation schemes. When the issue of time criticality is introduced, the sequence in which links are activated (in addition to the issue of the frequency with which a link is activated per unit time) and the order in which packets at a particular node are transmitted become important. Thus two distinct aspects to the scheduling problem must be addressed—the order in which links are activated and the sequence in which the packets at a particular node are transmitted. A natural question is whether or not acceptable performance can be achieved if the implementation of routing and scheduling (both of its aspects) is separated. In other words, should a node perform routing decisions independently of the packet extinction times (resulting in simpler algorithms) and then schedule the packet in each link optimally, or should both decisions be made concurrently (resulting, presumably, in better performance)? Thus we see that the introduction of time constraints further complicates the already complex problem of joint routing and scheduling, and will represent a challenging research area in the coming years.

Final Remarks

Communication under time constraints is one aspect of voice/data integration that we plan to study in the coming years. Scheduling problems such as these are generally quite complex. Thus far, Bhattacharya and Ephremides have successfully applied both dynamic programming and probabilistic techniques to the relatively simple versions of these problems. Such techniques hold promise for the more complex versions, but they may have to be aided and complemented by simulation and computation. Alternatively, the formulation of these problems as discrete event dynamic systems (DEDS), which are discussed in the next section, may provide a useful framework to address these problems.

11. DISCRETE EVENT DYNAMIC SYSTEM (DEDS) MODELS AND THEIR ROLE IN THIS STUDY

Communication networks display dynamic discontinuity in that their evolution in time depends on the complex interactions of the timing of various discrete events, such as the arrival or transmission of packets, as well as on the interactions induced by the protocols that control the processes in these systems. These characteristics suggest that it may be advantageous to model communication networks as discrete event dynamic systems (DEDS). This class of systems has received a great deal of attention from control theorists in the last few years. Important considerations in the modeling of DEDS include system specification, functional correctness, the effective usage of available resources (e.g. communication channels), the verification of orderly information flow, the maintenance of security, and more generally the effective management and control of these systems.

Unlike naturally occurring physical systems, which are continuous, DEDS are not easily described by standard tools such as differential or difference equations. Since DEDS have evolved only recently, their study does not have the benefit of a long history of mathematical modeling. Simulation, which can be prohibitively expensive while providing little insight, is often the only way to evaluate system performance. Therefore, fundamental research is needed to develop a framework, or paradigm, for the design and performance evaluation of such systems.

An important aspect of this study will be developing analytical techniques for the control and performance evaluation of integrated networks. In particular, we plan to develop new modeling techniques for such networks based on the DEDS framework. These studies will involve applying existing DEDS techniques to new networking models, as well as developing new techniques for the modeling of DEDS.

In this section we first outline the difficulties involved in developing a paradigm for DEDS. We then discuss some of the methods that have been used in the past to characterize such systems, as well as some promising new approaches that have created a lot of interest in the last few years. Finally, we discuss the application of DEDS models to integrated networks.

The DEDS Modeling Problem

The area of DEDS was identified as one of the important future research topics in control theory at a recent workshop (Levis et al. 1987). Subsequently, in an editorial Ho (1987a) discussed the importance of the development of a paradigm for DEDS while stressing the need to produce models that reflect real-world problems, rather than purely mathematical concepts. A recent special issue of the *Proceedings of the IEEE* was devoted to this area (Ho, ed. 1989). This is one of the best sources of material on DEDS, and several papers from this issue are referenced later in this section. A key observation that can be made after reading these papers is that, despite the efforts of a number of outstanding researchers, only modest progress has been made thus far.

The task of developing a comprehensive framework for the modeling of DEDS would be monumental. This is true because a complete model for DEDS would have to address the following issues:

- the discontinuous nature of discrete events;
- the continuous nature of most performance measures;
- the importance of probabilistic formulation;
- the need for modular analysis;
- the presence of dynamics;

- the feasibility of the computational burden; and

- the communicating processes within the system that govern its dynamical evolution.

This list has been adapted from Ho (1987a). The first and last items reflect the special characteristics of DEDS, as compared with continuous-variable dynamic systems (CVDS); the others are common to both discrete and continuous systems. Ultimately, the research community would like to have a model for DEDS that is as complete as existing models for CVDS. It is worth discussing two of these items at this point. First, modular analysis is desirable so that various aspects of system performance can be analyzed without requiring a complete system performance evaluation; a hierarchical analysis may be beneficial in some cases. Second, the computational burden is generally not prohibitive in CVDS because accurate models often exist; however, current approaches to DEDS often permit the evaluation of only trivially small examples. Parallel processing capabilities of modern machines may be useful in the evaluation of DEDS.

Existing Models for DEDS

The techniques that have been used to model DEDS are discussed by Ho (1987b). These include relatively conventional techniques that have been used for some time, as well as a variety of new approaches that are in early stages of development and show a great deal of promise.

Conventional Techniques

The most basic modeling technique for DEDS is the use of finite-state Markov chains or processes. The primary difficulty with such models is the combinatorial explosion in the number of states. For example, a communication network may be characterized by a vector consisting of the number of packets in queue at each node; an extremely large number of states result for all but the smallest systems. Furthermore, state transition probabilities must be evaluated for all pairs of states.

Another disadvantage of the use of finite-state Markov chain models is that the state description offers very little insight into the structural properties of the system being modeled. Since the state is characterized as a one-dimensional vector, relationships among states are difficult to exploit in developing dynamic system models. Petri nets (Peterson 1981) are a useful approach for graphically illustrating the relationships among the states in a dynamic system. However, they are best suited to answering qualitative questions about system performance, rather than being able to provide numerical performance measures. Another limitation is that they are generally useful only for relatively small systems.

Queueing network models (Kleinrock 1975; Gelenbe and Pujolle 1987; Walrand 1988) are useful for evaluating equilibrium behavior in systems that can be classified as "standard," in which case "product-form" solutions may be obtained. The major properties of standard systems include statistical independence of arrival and service processes, statistical independence of successive service times, and infinite buffer size. Although many queueing networks are quite robust with respect to these assumptions (Suri 1983) (i.e., product-form solutions often provide acceptable solutions even when these assumptions are violated, as they are in most practical systems), it is generally not possible to determine robustness without also doing an exact performance evaluation. Thus the applicability of standard queueing models is generally quite limited. "Operational analyses" (Denning and Buzen

1978; Buzen and Denning 1980) that are based on measured quantities rather than probabilistic properties have been developed that do not require that the standard properties be satisfied. However, even these approaches run into difficulty in systems with finite queue limit, state-dependent routing, simultaneous resource sharing, and nonstandard queue disciplines. A fundamental limitation of all of these queueing network approaches is that they are only quasi-dynamic in nature, i.e., although system dynamics are incorporated into the model, system performance can be evaluated only on an average basis in the steady state. Thus, these techniques cannot be used for transient analysis, and an adequate framework for incorporating control mechanisms does not exist. Limitations on the applicability of the queueing models to communication networks are discussed by Ephremides (1986).

It was noted earlier that simulation can be a useful tool for the performance evaluation of DEDS but that its effectiveness is often limited by cost and time considerations. Research in simulation methods has emphasized the development of good simulation languages to facilitate modeling. Also. statistical analysis has been used to make simulation experiments more efficient. For example, probability transformations have been used to make "unlikely" events more probable, thus shortening the time required for simulation runs (Parekh and Walrand 1989; Cottrell et al. 1983). In addition, mathematical techniques that enhance the simulation process have been studied. For example, under the theory of generalized semi-Markov processes (GSMPs), one distinguishes between the discrete/countable and continuous/uncountable parts of the state space (Whitt 1980; Schassberger 1976; Glynn 1989). Simulation languages have been developed by using this approach. But a complete mathematical framework to support such efforts has not yet been developed.

New Techniques

Recently, a number of new techniques have been applied to the modeling and analysis of DEDS. The first that we consider, which is basically a control theory approach, is perturbation analysis (PA). This technique was developed primarily by Ho and his associates (Ho 1985, 1987b, 1988; Ho and Li 1988; Gong and Ho 1987; Cassandras and Strickland 1989) and by Suri (1987, 1989). Under this approach, system dynamics are "linearized" about a particular trajectory. A single simulation run is executed, and the sensitivity of system performance to various system parameters is analyzed, thereby permitting one to obtain performance results without running a separate simulation for each set of parameters.

Figure 4 shows the principles of PA. The network dynamics are driven by a vector of random variables $\omega(k;\theta)$ where k represents time and θ is a specified parameter vector. The performance measure for a specific realization is denoted as $J(w;\theta)$. The objective of PA is to estimate the sample path sensitivity $\Delta J(w;\Delta\theta)$, given a perturbation $\Delta\theta$, when $J(w;\theta)$ is observed. Although used in conjunction with simulation, PA is primarily an analytical approach that provides more information from a limited number of simulations.

The ability of PA techniques to provide performance sensitivities with respect to given parameter perturbations permits one to obtain the answers to "what if ..." questions without executing many simulation runs. In addition, PA permits the estimation of performance gradients with respect to system parameters. Such performance gradients are normally required for the execution of stochastic optimization algorithms. Important considerations relating to the applicability of PA include the size of perturbations that can be handled without requiring additional simulation runs, and the related issue of whether such methods can be applicable when perturbations result in changes in the order in which events occur in the system.

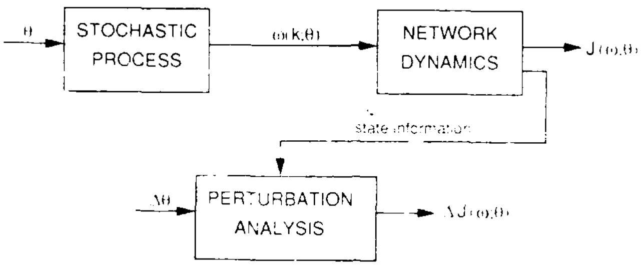

Fig. 4 — Perturbation analysis framework

PA has been applied successfully to a number of queueing and networking problems (Ho and Cao 1983, 1985; Cao and Ho 1987; Cao 1987, 1989; Cassandras and Strickland, 1988; Cassandras et al. 1988; Suri 1989; Vakili and Ho 1987), but considerably more work is needed. In particular, Cassandras and Strickland have applied PA to distributed routing. In this problem the derivatives of delay with respect to link flows are needed for the minimization of delay with respect to link flows, and PA is used to determine these derivatives.

The modeling of DEDS can also be approached from the perspective of computer science in the form of "language development" for supervisory control of DEDS, an approach developed by Wonham and Ramadge (Ramadge and Wonham 1987, 1989; Wonham and Ramadge 1987, 1988; Ramadge 1989). The problem addressed by such approaches is that of achieving or verifying the orderly flow of events in a dynamic system. The techniques used are derived from logic, language, and automata theory (see e.g., Hopcroft and Ullman 1979). A related approach is that of using finitely recursive processes to model DEDS, a technique developed by Inan and Varaiya (1988, 1989), based on Hoare's theory of communicating sequential processes (Hoare 1985). Perhaps the most obvious application of either of these techniques is the verification of protocol structures (i.e., the orderly flow of events in a network). However, advances in these areas may permit the development of comprehensive control models that permit the implementation of system controllers and the evaluation of system performance as well. Another approach is the development of algebraic tools to characterize a certain class of DEDS that can be modeled by timed event graphs, which are a special class of Petri nets (Cohen et al. 1989).

Application of DEDS Techniques to Voice/Data Integration

The DEDS approach appears to be a natural framework in which to study integrated voice/data networks. To illustrate this, we first consider operation at a single node. Each node in a network can be characterized in terms of a state (m,n), where m is the number of calls in progress and n is the number of data packets in the buffer. The control action taken by a node is the acceptance or blockage of a voice call. Clearly there is a trade-off, since the acceptance of voice calls results in less channel capacity being available for data traffic, and hence an increase in queue size, which in turn results in increased delay. As noted earlier, the control problem at a node can be formulated in any of three basic ways:

- minimize packet delay subject to a constraint on blocking probability,
- minimize blocking probability subject to a constraint on packet delay, or
- minimize a weighted sum of blocking probability and packet delay.

Most studies of voice/data integration systems have considered operation at only a single node. An overall system description would be characterized by a composite state that contains the state of each node, as discussed here. Events include message arrivals and successful transmissions. System dynamics are driven by the complex interactions of such events and must reflect the characteristics of network traffic (e.g., arrival statistics and message lengths). Global performance criteria (e.g. network throughput) must be established, in contrast to the purely local criteria discussed here. Control schemes must be developed under which the control actions taken by each node are based on its own state, and possibly on its estimates of the states of other nodes in the network. Such estimates may be derived from observations of network traffic as well as from the information explicitly transmitted in status messages. Furthermore, such control actions are based on the performance criteria that have been chosen for the particular network application.

Final Remarks

In addition to using DEDS techniques, we will also use other analytical techniques for performance evaluation and optimization. For example, in Appendixes A, B, and C we discuss the formulation of the movable-boundary problem as a Markovian decision process problem, and we indicate how several techniques may be used to solve such problems. However, the applicability of these methods may be limited to operation at a single node, or possibly extended to simple network configurations such as tandem networks. New methods are needed to permit the analysis of more general network models. We feel that DEDS may provide the framework that is necessary for this task.

In conclusion, we believe that the study of DEDS will provide a fruitful framework for the study of integrated voice/data networks. Our study will involve the development of new DEDS techniques, as well as the application of both new and existing techniques to the modeling and control of such networks. We note that although simulation will be used in these studies, the emphasis will be on the development and application of analytical techniques, including those that may be used in conjunction with the simulation process to enhance its effectiveness.

12. SUMMARY, CONCLUSIONS, AND FUTURE PLANS

The transport of both voice and data will be an important requirement of many future military communication networks. In this report we have discussed many of the fundamental issues associated with the design of such integrated networks, and we have indicated areas that require future basic research. This report represents our preliminary assessment of the area of voice/data integration after a short period of background investigation. Thus many of our conclusions are necessarily tentative at this point; they should be interpreted as the basis for future research plans, rather than as an assessment of ultimate network designs. In this concluding section, we review the voice/data integration problem briefly, and we outline our future research plans in this area.

Most studies of voice/data integration have investigated wireline or satellite networks, which are characterized by contention-free links. In this study we are interested primarily in mobile radio networks. The nature of these networks creates a totally new networking environment in which the problems of voice/data integration have scarcely been addressed. In particular, the broadcast property of radio networks results in a much greater degree of interaction among network nodes than in wireline or satellite networks. Also, mobile radio networks are characterized by a dynamic topology, requiring the development of network control schemes that are capable of adapting rapidly and robustly to topological change. Although a variety of such schemes has been developed for data networks, considerable work is still needed. When the additional features of integrated networks must be considered, we see that this area is still in its infancy. Furthermore, when the further complications that arise from considerations related to operation in a hostile Navy environment are considered (e.g., jamming, platform destruction, security issues, and precedence requirements), it is easily recognized that this is a critical area in which basic research must be pursued before survivable, robust, and efficient network designs can be developed.

In this report we have indicated how voice and data traffic impose different and often conflicting demands on network operation. For example, voice traffic requires nearly real-time delivery, while data traffic can usually tolerate some delay. Another difference is that voice traffic can tolerate relatively high error rates, while data traffic cannot.

We reviewed the switching methods that can be used in networks. Generally, packet switching is preferable for bursty data, while circuit switching or virtual circuit switching (actually a form of packet switching) is generally preferable for voice. Thus a hybrid form of switching that combines features of both packet and (virtual) circuit switching is often used in integrated networks. Although such an approach appears reasonable for mobile military networks, it is still too soon to make definite conclusions on a switching method. A crucial consideration is whether it will be possible to implement virtual circuit switching in a network with rapidly changing topology. To achieve the required levels of adaptability and robustness, it may be necessary to develop purely packet-switched (datagram) schemes. This is an important subject for future research.

Voice/data integration efforts in the commercial sector have concentrated on the development of the Integrated Services Digital Network (ISDN), a network that will be able to support a wide variety of digital communication traffic. These studies have emphasized the development of standards for the ISDN, rather than the development of exact analytical models for such networks. To put our studies in perspective, we have reviewed some of the major characteristics of the ISDN.

Before addressing methods for actual integration of voice and data, we reviewed speech interpolation methods that permit more efficient channel use by exploiting the characteristics of speech traffic. Critical issues include whether or not any form of speech interpolation will be possible with the low data rate channels used in military applications, and whether or not the buffering of speech will facilitate speech interpolation.

Research efforts in the literature have concentrated on the movable-boundary method for multiplexing data and voice. However, very little has been accomplished in the area of optimization of such schemes. Furthermore, analysis has generally been limited to the case of a single isolated node because of the complexity of modeling multihop operation. Thus, even after a decade or so of study, the area of voice/data integration remains in need of additional basic research. We reviewed some of the major variations of the movable-boundary scheme and the most significant studies to date. We then discussed some of the issues associated with networkwide control and channel access in integrated networks. The extension from a single node to a tandem node configuration results in considerable complication, and few results have been obtained in this area. Virtually nothing has been accomplished in the area of control of more general network configurations. As a step in this direction, we have reviewed some of the issues associated with communication under time constraints. This area is important to voice traffic and to time-critical data traffic.

A variety of methods have been used to evaluate integrated networks. Except for the simplest of configurations, the use of approximations or simulation has generally been necessary. In Appendixes A, B, and C we discuss the formulation of the optimal control of the movable-boundary scheme as a Markovian decision process problem. This approach has permitted some extension to the case of tandem networks.

However, it is clear that new approaches are needed if we are to make real breakthroughs in the study of integrated networks. We feel that a promising approach is the use of discrete event dynamic systems (DEDS) models to characterize network operation. We have reviewed the basic characteristics of DEDS, and we have indicated why we feel that these techniques may provide the framework that is necessary to model many communication networks and other problems of interest to the Navy.

Future Plans

We are now in the early stages of a basic research study entitled "Modeling Techniques for DEDS and Voice/Data Integration in Radio Networks." Our plans for the next 5 years are summarized as follows:

In FY89 we will study existing models for DEDS and examine their applicability to problems of Navy interest, which are not necessarily limited to the voice/data integration problem. After such preliminary studies, we will identify new directions for basic research in DEDS that reflect Navy needs and begin research in one or more of the most promising areas. At the same time, we will continue our study of existing techniques for voice/data integration and methods currently available for their evaluation and optimization. We will develop preliminary networking models that exploit the DEDS framework, with emphasis on integrated voice/data radio networks. Alternative approaches for analysis and control, such as the use of Markovian decision process models, will also be considered, and their possible relationships to DEDS models will be studied. At this stage of our studies, the emphasis will be on small and simple networks because of the difficulty of modeling more complex configurations.

In FY90 we will develop new DEDS techniques to permit evaluation of the networking models developed in the previous year. We will also develop improved models and control techniques for integrated networks. The accuracy of the models developed in the previous year for small networks will be improved. Also, models for larger and more complicated networks will be developed. The performance of integrated voice/data networks will be evaluated by analysis where possible and by simulation where necessary. Where simulation is used, the emphasis will be on analytical techniques that enhance the effectiveness of the simulation, rather than on the development of simulation techniques themselves.

In FY91 we will continue basic research in techniques for the modeling, control, and evaluation of DEDS. In addition to voice/data integration, we will identify other areas for application of DEDS techniques, where appropriate. We will continue to develop improved models and control techniques for voice/data integration in radio networks. Our studies to date will be integrated into a preliminary system concept for integrated voice/data radio networks that addresses channel access, routing, flow control, and network organization.

In FY92 we will continue to develop DEDS techniques for use with integrated voice/data radio networks and other Navy applications. We will also continue to improve integrated voice/data radio network models and control schemes.

In FY93 we will develop a system concept for integrated voice/data radio networks that addresses all aspects of network design. As we complete this study, we plan to transition to a 6.2 task in voice/data integration and possibly other areas that can exploit DEDS models.

REFERENCES

Arthurs, E. and B.W. Stuck (1983), "Traffic Analysis Tools for Integrated Digital Time-Division Link Level Multiplexing of Synchronous and Asynchronous Message Streams," *IEEE J. Sel. Areas Commun.* **SAC-1**, 1112-1123.

Baker, D.J., A. Ephremides, and J.E. Wieselthier (1982), "A Distributed Algorithm for Scheduling the Activation of Links in a Self-Organizing Mobile Radio Network," *Conf. Rec. 1982 Int. Conf. Commun.*, pp. 2F.6.1-2F.6.5.

Baker, D.J., A. Ephremides, and J.A. Flynn (1984), "The Design and Simulation of a Mobile Radio Network with Distributed Control," *IEEE J. Sel. Areas Commun.* **SAC-2**, 226-237.

Bhattacharya, P.P. and A. Ephremides (1988), "Optimal Scheduling of the Transmission of Messages with Strict Deadlines," *Proc. 22nd Conf. Info. Sciences Syst. (CISS)*, Princeton University, pp. 623-628.

Bhattacharya, P.P. and A. Ephremides (1989), "Optimal Scheduling with Strict Deadlines," *IEEE Trans. Auto. Control* **34,** 721-728.

Bially, T., B. Gold, and S. Seneff (1980a), "A Technique for Adaptive Voice Flow Control in Integrated Packet Networks," *IEEE Trans. Commun.* **COM-28**, 325-333.

Bially, T., A.J. McLaughlin, and C.J. Weinstein (1980b), "Voice Communication in Integrated Digital Voice and Data Networks," *IEEE Trans. Commun.* **COM-28**, 1478-1490.

Bolger, T.E. (1988), "The Controversies Surrounding ISDN," *Comp. Networks ISDN Syst.* **15**, 27-30.

Brady, P.T. (1968), "A Statistical Analysis of On-Off Patterns in 16 Conversations," *Bell Sys. Tech. J.* **47**, 73-91.

Bullington, K. and J.M. Fraser (1982), "Engineering Aspects of TASI," *Bell Syst. Tech. J.* **38**.

Buzen, J.P. and P.J. Denning (1980), "Measuring and Calculating Queue Length Distributions," *Computer* **13**, 33-44.

Campanella, S.J. (1987), "Digital Speech Interpolation Systems," in *Advanced Digital Communications: Systems and Signal Processing Techniques*, ed. K. Feher (Prentice-Hall, New York, NY), pp. 237-281.

Cao, X.-R. (1987), "The Static Property of Perturbed Queuing Networks," *Proc. 26th Conf. Decision Control (CDC)*, pp. 257-262.

Cao, X.-R. and Y.-C. Ho (1987), "Sensitivity Analysis and Optimization of Throughput in a Production Line with Blocking," *IEEE Trans. Auto. Control* **AC-32**, 959-967.

Cao, X.-R. (1989), "A Comparison of the Dynamics of Continuous and Discrete Event Systems" *Proc. IEEE* **77**, 7-13.

Cassandras, C., M.V. Abidi, and D. Towsley (1988), "Distributed Routing with On-Line Marginal Delay Estimation," *Conf. Rec. INFOCOM '88*, pp. 603-612.

Cassandras, C.G. and S.G. Strickland (1988), "Perturbation Analytic Methodologies for Design and Optimization of Communication Networks," *IEEE J. Sel. Areas Commun.* **6**, 158-171.

Cassandras, C.G. and S.G. Strickland (1989), "Sample Path Properties of Timed Discrete Event Systems," *Proc. IEEE* **77**, 59-71.

Chu, W.W. (1969), "A Study of Asynchronous Time Division Multiplexing for Time-Sharing Computer Systems," *Proc. Fall Joint Computing Conf.*

Cohen, G., P. Moller, J.-P. Quadrat, and M. Viot (1989), "Algebraic Tools for the Performance Evaluation of Discrete Event Systems," *Proc. IEEE* **77**, 39-58.

Cottrell, M., J.-C. Fort, and G. Malgouyres (1983), "Large Deviations and Rare Events in the Study of Stochastic Algorithms," *IEEE Trans. Auto. Control* **AC-28**, 907-920.

Coviello, G.J. and P.A. Vena (1975), "Integration of Circuit/Packet Switching by a SENET (Slotted Envelope Network) Concept," *Conf. Rec. Nat. Telecommun. Conf.*, pp. 42.12-42.17.

Coviello, G.J. (1979), "Comparative Discussion of Circuit- vs. Packet-Switched Voice," *IEEE Trans. Commun.* **COM-27**, 1153-1160.

Daigle, J.N. and J.D. Langford (1986), "Models for Analysis of Packet Voice Communications Systems," *IEEE J. Sel. Areas Commun.* **SAC-4**, 847-855.

Decina, M. and D. Vlack, ed. (1983) "Special issue "Packet Switched Voice and Data Communications," *IEEE J. Sel. Areas Commun.* **SAC-1**.

Decina, M., W.S. Gifford, R. Potter, and A.A. Robrock, eds. (1986), Special issue "Integrated Services Digital Network: Recommendations and Field Trials—I," *IEEE J. Sel. Areas Commun.* **SAC-4**.

Decina, M., W.S. Gifford, R. Potter, and A.A. Robrock, eds. (1986), Special issue "Integrated Services Digital Network: Technology and Implementation—II," *IEEE J. Sel. Areas Commun.* **SAC-4**.

Decina, M. and A. Roveri (1987), "ISDN: Integrated Services Digital Network: Architectures and Protocols," in *Advanced Digital Communications: Systems and Signal Processing Techniques*, ed. K. Feher (Prentice-Hall, New York, N.Y.), pp. 40-132.

Denning, P.J. and J.P. Buzen (1978) "The Operational Analysis of Queueing Network Models," *Computing Surveys*, **10**, 225-261.

Douligeris, C. and R. Mazumdar (1988), "A Game Theoretic Approach to Flow Control in an Integrated Environment with Two Classes of Users," *Conf. Rec. Computer Networking Symp.*, pp. 214-221.

Ephremides, A. (1988), "Limitations of Queueing Models in Communication Networks," *Proc. of the NATO Advanced Study Inst. on Performance Limits Commun. Theory Practice*, ed. J.K. Skwirzynski, pp. 143-154.

Fischer, M.J. and T.C. Harris (1976), "A Model for Evaluating the Performance of an Integrated Circuit- and Packet-Switched Multiplex Structure," *IEEE Trans. Commun.* **COM-24**, 195-202.

Fraser, J.M., D.B. Bullock, and H.G. Long (1962), "Overall Characteristics of a TASI System," *Bell Syst. Tech. J.* **41**.

Gallager, R.G. (1977), "A Minimum Delay Routing Algorithm Using Distributed Computation," *IEEE Trans. Commun.* **COM-25**, 73-85.

Gaver, D.P. and J.P. Lehoczky (1982), "Channels That Cooperatively Service a Data Stream and Voice Messages," *IEEE Trans. Commun.* **COM-30**, 1153-1162.

Gelenbe, E. and G. Pujolle (1987), *Introduction to Queueing Networks* (John Wiley and Sons, New York, NY).

Geraniotis, E. and J.W. Gluck (1987), "Coded FH/SS Communications in the Presence of Combined Partial-Band Noise Jamming, Rician Nonselective Fading, and Multiuser Interference," *IEEE J. Sel. Areas Commun.* **SAC-5**, 194-214.

Gerla, M. (1985), "Packet, Circuit, and Virtual Circuit Switching," in *Computer Communications—Vol. II: Systems and Applications*, W. Chou, ed. (Prentice-Hall, New York, NY), pp. 222-267.

Gitman, I. and H. Frank (1978), "Economic Analysis of Integrated Voice and Data Networks: A Case Study," *Proc. IEEE* **66** 1549-1570.

Glynn, P.W. (1989), "A GSMP Formalism for Discrete Event Systems," *Proc. IEEE* **77**, 14-23.

Gold, B. (1977) "Digital Speech Networks," *Proc. IEEE* **65**, 1636-1658.

Gong, W.-B. and Y.-C. Ho (1987), "Smoothed (Conditional) Perturbation Analysis of Discrete Event Dynamical Systems," *IEEE Trans. Auto. Control* **AC-32**, 858-866.

Gruber, J.G. (1981), "Delay Related Issues in Integrated Voice and Data Networks," *IEEE Trans. Commun.* **COM-29**, 786-800.

Gruber, J.G. and N.H. Le (1983), "Performance Requirements for Integrated Voice/Data Networks," *IEEE J. Sel. Areas Commun.* **SAC-1**, 981-1005.

Harrington, E. A. (1980), "Voice/Data Integration Using Circuit Switched Networks," *IEEE Trans. Commun.* **COM-28**, 781-793.

Ho Y.-C. and X. Cao (1983), "Perturbation Analysis and Optimization of Queueing Networks," *J. Optim. Theory Applic.* **40**, 559-582.

Ho, Y.-C. (1985), "On the Perturbation Analysis of Discrete-Event Dynamic Systems," *J. Optim. Theory Applic.* **46**, 535-545.

Ho, Y.-C. and X.-R. Cao (1985), "Performance Sensitivity to Routing Changes in Queueing Networks and Flexible Manufacturing Systems Using Perturbation Analysis," *IEEE J. Robotics Automation* **RA-1**, 165-172.

Ho, Y.-C. (1987a) "Editorial: Basic Research, Manufacturing Automation, and Putting the Cart Before the Horse," *IEEE Trans. Auto. Control* **AC-32**, 1042-1043.

Ho, Y.-C. (1987b) "Performance Evaluation and Perturbation Analysis of Discrete Event Dynamic Systems," *IEEE Trans Auto. Control* **AC-32**, 563-572.

Ho, Y.-C. and S. Li (1988), "Extensions of Infinitesimal Perturbation Analysis." *IEEE Trans. Auto. Control* **33**, 427-438.

Ho, Y.-C. (1988), "Perturbation Analysis Explained," *IEEE Trans. Auto. Control* **33**, 761-763.

Ho, Y.-C., ed. (1989), Special issue on "Dynamics of Discrete Event Systems," *Proc. IEEE* **77**.

Hoare, C.A.R. (1985), *Communicating Sequential Processes* (Prentice-Hall, New York, NY).

Hoberecht, W.L. (1983), "A Layered Network Protocol for Packet Voice and Data Integration," *IEEE J. Sel. Areas Commun.* **SAC-1**, 1006-1013.

Hopcroft, J.E. and J.D. Ullman (1979) *Introduction to Automata Theory, Languages, and Computation* (Addison-Wesley, Reading, MA).

Ibe, O.C. (1985), "Modeling of Integrated Packet-Switched Voice and Data Networks," *Conf. Rec. IEEE GLOBECOM'85*, pp. 739-745.

Inan, K. and P. Varaiya (1988), "Finitely Recursive Process Models for Discrete Event Systems," *IEEE Trans. Auto. Control* **33**, 626-639.

Inan, K.M. and P.P. Varaiya (1989), "Algebras of Discrete Event Models," *Proc. IEEE* **77**, 24-38.

Janakiraman, N., B. Pagurek, and J.E. Neilson (1984a), "Performance Analysis of an Integrated Switch with Fixed or Variable Frame Rate and Movable Voice/Data Boundary," *IEEE Trans. Commun.* **COM-32**, 34-39.

Janakiraman, N., B. Pagurek, and J.E. Neilson (1984b), "Delay Versus TASI Advantage in a Packet Voice Multiplexer," *IEEE Trans. Commun.* **COM-32**, 319-320.

Kang, G.S. and L.J. Fransen (1982), "Second Report of the Multirate Processor (MRP) for Digital Voice Communications," NRL Report 8614.

Kang, G.S. and L.J. Fransen (1985), "Low-Bit Rate Speech Encoders Based on Line-Spectrum Frequencies (LSFs)," NRL Report 8857.

Katevenis, M.G.H. (1987), "Fast Switching and Fair Control of Congested Flow in Broadband Networks," *IEEE J. Sel. Areas Commun.* **SAC-5**, 1315-1326.

Kleinrock, L. (1975), *Queueing Systems, Vol. I: Theory* (John Wiley & Sons, New York, NY).

Konheim, A.G. and R.L. Pickholtz (1984), "Analysis of Integrated Voice/Data Multiplexing," *IEEE Trans. Commun.* **COM-32**, 140-147.

Kum: ıerle, K. (1974), "Multiplexer Performance for Integrated Line- and Packet-Switched Traffic," *Proc. Second Int. Conf. Computer Commun. (ICCC)*, pp. 508-515.

Kurose, J.F., M. Schwartz, and Y. Yemini (1984), "Multiple-Access Protocols and Time-Constrained Communication," *Computing Surveys* **16**, 43-70.

Kurose, J.F., M. Schwartz, and Y. Yemini (1988), "Controlling Window Protocols for Time-Constrained Communication in Multiple Access Networks," *IEEE Trans. Commun.* **36**, 41-49.

Lam, S.S. (1980), "Packet Broadcast Networks—A Performance Analysis of the R-ALOHA Protocol," *IEEE Trans. Computers* **C-29**, 596-603.

Lam, S.S. (1983) "Multiple Access Protocols," in *Computer Communications Vol. I: Principles*, ed. W. Chou (Prentice-Hall, New York, NY).

Lambadaris, I., P. Narayan, and I. Viniotis (1987), "Optimal Service Allocation Among Two Homogeneous Traffic Types with No Queueing," *Proc. 26th Conf. Decision Control (CDC)*, pp. 1496-1498.

Lee, H.H. and C.K. Un (1985), "Performance Analysis of Statistical Voice/Data Multiplexing Systems with Voice Storage," *IEEE Trans. Commun.* **COM-33**, 809-819.

Leon-Garcia, A., R.H. Kwong, and G.F. Williams (1982), "Performance Evaluation Methods for an Integrated Voice/Data Link," *IEEE Trans. Commun.* **COM-30**, 1848-1858.

Levis, A. et al. (1987), "Challenges to Control: A Collective View, Report of the Workshop Held at the University of Santa Clara on September 18-19, 1986," *IEEE Trans. Auto. Control* **AC-32**, 275-285.

Li, S.-q. and J.W. Mark (1984), "Integrated Services on a TDM with Window Flow Control," *Conf. Info. Sciences Syst. (CISS)*, Princeton University, pp. 562-566.

Lippman, R.P. (1985), "New Routing and Preemption Algorithms for Circuit-Switched Mixed-Media Networks," *Conf. Rec. IEEE MILCOM '85*, pp. 660-666.

Maglaris, B.S. and M. Schwartz (1982), "Optimal Fixed Frame Multiplexing in Integrated Line- and Packet-Switched Communication Networks," *IEEE Trans. Inf. Theory* **IT-28**, 263-273.

Maglaris, B. and M. Schwartz (1981), "Performance Evaluation of a Variable Frame Multiplexer for Integrated Switched Networks," *IEEE Trans. Commun.* **COM-29**, 800-807.

Maglaris, B., R. Boorstyn, S. Panwar, and T. Spirtos (1987), "Routing in Burst-Switched Voice/Data Integrated Networks," *Conf. Rec. IEEE INFOCOM'87*, pp. 162-169.

Mercer, R.A. and W.L. Edwards (1986), "Issues in the Migration of Military Communications to an ISDN," *Conf. Rec. IEEE MILCOM'86*, 50.5.1-50.5.5.

Midema, H. and M.G. Schachtman (1962), "TASI Quality-Effect of Speech Detectors and Interpolation," *Bell Sys. Tech. J.* **41**.

Miyahara, H. and T. Hasegawa (1978), "Integrated Switching with Variable Frame and Packet," *Conf. Rec. Int. Conf. Commun. (ICC78)*, pp. 20.3.1-20.3.5.

Panwar, S.S., D. Towley, J.K. Wolf, and Y. Armoni (1987), "Collision Resolution Algorithms for a Time-Constrained Multiaccess Channel," *Proc. Twenty-Fifth Annual Allerton Conf. Commun., Control, and Computing*, pp. 1081-1088.

Parekh, S. and J. Walrand (1989), "A Quick Simulation Method for Excessive Backlogs in Networks of Queues," *IEEE Trans. Auto. Control* **34**, 54-66.

Paterakis, M., L. Georgiadis, and P. Papantoni-Kazakos (1987), "Multiple Access Policies for Systems with Strict Delay Limitations," *Proc. Command Control Res. Symp.*, National Defense University, Fort Lesley J. McNair, pp. 177-182.

Peterson, J.L. (1981), *Petri Net Theory and the Modeling of Systems* (Prentice-Hall, New York, NY).

Pickens, R.A. and K.W. Hanson (1985) "Integrating Data, Voice, and Image," in *Computer Communications Vol. II: Systems and Applications*, W. Chou, ed. (Prentice-Hall, New York, NY), pp. 268-356.

Ramadge, P.J. and W.M. Wonham (1987), "Supervisory Control of a Class of Discrete Event Processes," *SIAM J. Control Optim.* **25**, 206-230.

Ramadge, P.J.G. (1989), "Some Tractable Supervisory Control Problems for Discrete-Event Systems Modeled by Buchi Automata," *IEEE Trans. Auto. Control* **34**, 10-19.

Ramadge, P.J.G. and W.M. Wonham (1989), "The Control of Discrete Event Systems," *Proc. IEEE* **77**, 81-98.

Roberts, L.G. (1973), "Dynamic Allocation of Satellite Capacity Through Packet Reservation," *AFIPS Conf. Proc., 1973 Nat. Computer Conf.* **42**, pp. 711-716.

Roehr, W. (1987), "ISDN—Is It Enough to Satisfy Military Needs?," *Conf. Rec. IEEE MILCOM'87*, pp. 813-837.

Ronayne, J. (1988), *The Integrated Services Digital Network: from Concept to Application* (John Wiley and Sons, New York, NY).

Ross, K.W. and D. Tsang (1989), "Optimal Circuit Access Policies in an ISDN Environment: A Markov Decision Approach," *IEEE Trans. Commun.* **37**, 934-939.

Ross, M.J. and O.A. Mowafi (1982), "Performance Analysis of Hybrid Switching Concepts for Integrated Voice/Data Communications," *IEEE Trans. Commun.* **COM-30**, 1073-1087.

Schassberger, R. (1976), "On the Equilibrium Distribution of a Class of Finite-State Generalized Semi-Markov Processes," *Math. Operations Res.* **1**, 395-406.

Shacham, N., E.J. Craighill, and A.A. Poggio (1983), "Speech Transport in Packet-Radio Networks with Mobile Nodes," *IEEE J. Sel. Areas Commun.* **SAC-1**, 1084-1097.

Silverman, S. (1987), "ISDN and the DOD," *Conf. Rec. MILCOM'87*, pp. 775-778.

Sriram, K., P. Varshney, and J.G. Shanthikumar (1983), "Discrete-Time Analysis of Integrated Voice/Data Multiplexers With and Without Speech Activity Detectors," *IEEE J. Sel. Areas Commun.* **SAC-1**, 1124-1132.

Stallings, W., ed. (1985), *Tutorial: Integrated Services Digital Networks (ISDN)*, (IEEE Computer Society Press, New York, NY).

Suda, T., H. Miyahara, and T. Hasegawa (1983), "Performance Evaluation of an Integrated Access Scheme in a Satellite Communication Channel," *IEEE J. Sel. Areas Commun.* **SAC-1**, 153-164.

Suri, R. (1983), "Robustness of Queueing Network Formulas," *J. Assoc. Computing Machinery* **30**, 564-594.

Suri, R. (1987), "Infinitesimal Perturbation Analysis for General Discrete Event Systems," *J. Assoc. Computing Machinery* **34**, 686-717.

Suri, R. (1989), "Perturbation Analysis: The State of the Art and Research Issues Explained via the G/G/1 Queue," *Proc. IEEE* **77**, 114-137.

Tanenbaum, A.S. (1988), *Computer Networks, Second Edition* (Prentice Hall, New York, NY).

Tassiulas, L. and A. Ephremides (1989), "An Algorithm for Joint Routing and Scheduling in Radio Networks," Conf. Rec. American Control Conference.

Thomopoulos, S.C.A. and L. Zhang (1987), "Flow Control in ISDN with Limited Buffering and Discouragement," *Conf. Rec. 28th Conf. Decision Control (CDC)*, 697-702.

Tijms, H.C. (1986), *Stochastic Modelling and Analysis: A Computational Approach* (John Wiley & Sons, New York, NY).

Tobagi, F.A. (1980), "Multiaccess Protocols in Packet Communication Systems," *IEEE Trans. Commun.* **COM-28** 468-488.

Tsang, D. and K.W. Ross (1988), "Structural Properties of Optimal Policies in an ISDN Circuit-Switched Access Port," *Conf. Rec. IEEE Computer Networking Symp.*, pp. 2-7.

Upton, R.A. (1984), "A Transient Analysis of Computer Communication Networks," Ph.D. dissertation, University of Maryland.

Vakili, P. and Y.-C. Ho (1987), "Infinitesimal Perturbation Analysis of a Multiclass Routing Problem," *Conf. Rec. 25th Ann. Allerton Conf. Commun., Control, Computing*, pp. 279-286.

Viniotis, I. and A. Ephremides (1987), "Optimal Switching of Voice and Data at a Network Node," *Proc. 28th Conf. Decision Control (CDC)*, pp. 1504-1507.

Viniotis, I. (1988), "Optimal Control of Integrated Communication Systems Via Linear Programming Techniques," Ph.D. dissertation, University of Maryland.

Viniotis, I. and A. Ephremides (1988a), "On the Optimal Dynamic Switching of Voice and Data in Communication Networks," *Proc. Computer Networking Symp.*, pp. 8-16.

Viniotis, I. and A. Ephremides (1988b), "Linear Programming as a Technique for Optimization of Queueing Systems," *Proc. 27th Conf. Decision Control*, pp. 652-656.

Walrand J. (1984), ''A Note on Optimal Control of a Queueing System with Two Heterogeneous Servers,'' *Syst. Control Lett.* **4**, 131-134.

Walrand, J. (1988), *An Introduction to Queueing Networks*, (Prentice Hall, New York, NY).

Weinstein, C.J. and E.M. Hofstetter (1979), ''The Tradeoff Between Delay and TASI Advantage in a Packetized Speech Multiplexer,'' *IEEE Trans. Commun.* **COM-27**, 1716-1720.

Weinstein, C.J., M.L. Malpass, and M.J. Fischer (1980), ''Data Traffic Performance of an Integrated Circuit- and Packet-Switched Multiplex Structure,'' *IEEE Trans. Commun.* **COM-28**, 873-878.

Weinstein, C.J. and J.W. Forgie (1983), ''Experience with Speech Communication in Packet Networks,'' *IEEE J. Sel. Areas Commun.* **SAC-1**, 963-980.

Whitt, W. (1980), ''Continuity of Generalized Semi-Markov Processes,'' *Math. Oper. Res.* **5**, 494-501.

Wieselthier, J.E. and A. Ephremides (1980), ''A New Class of Protocols for Multiple Access in Satellite Networks,'' *IEEE Trans. Auto. Control* **AC-25**, 865-879.

Wieselthier, J.E. (1988), ''Code-Division Multiple-Access Techniques and Their Application to the High-Frequency (HF) Intratask Force (ITF) Communication Network,'' NRL Report 9094.

Wonham, W.M. and P.J. Ramadge (1987), ''On the Supremal Controllable Sublanguage of a Given Language,'' *SIAM J. Control Optim.* **25**, 637-659.

Wonham, W.M. and P.J. Ramadge (1988), ''Modular Supervisory Control of Discrete-Event Systems,'' *Math. Control, Signals, Syst.* **1**, 13-30.

Wu, C.-S. and V.O.K. Li (1988), ''Integrated Voice/Data Protocols for Satellite Channels.'' *Conf. Rec. Mobile Satellite Conf.*, Jet Propulsion Lab., pp. 413-422.

Yin, N., T.E. Stern, and S.-q. Li (1987a), ''Performance Analysis of a Priority-Oriented Packet Voice System,'' *Conf. Rec. IEEE INFOCOM'87*, pp. 856-863.

Yin, N., S.-q. Li, and T.E. Stern (1987b), ''Congestion Control for Packet Voice by Selective Packet Discarding,'' *Conf. Rec. IEEE GLOBECOM'87*, pp. 1782-1786.

Yum, T.-K. and M. Schwartz (1987), ''Packet-Switched Performance with Different Circuit-Switched Routing Procedures in Nonhierarchical Integrated Circuit-Switched and Packet-Switched Networks,'' *IEEE Trans. Commun.* **COM-35**, 362-366.

Zafiropoulo, P. (1974), ''Flexible Multiplexing for Networks Supporting Line-Switched and Packet-Switched Data Traffic,'' *Proc. Second Int. Conf. Computer Commun. (ICCC)*, pp. 517-523.

Appendix A

MARKOVIAN DECISION PROCESS PROBLEMS AND A GENERIC OPTIMIZATION APPROACH

A major goal of our studies is the development of control schemes that will provide optimal or near-optimal performance in integrated networks. We have noted that even the problem of multiplexing at a single node (e.g., using the movable-boundary scheme) has been a rather poorly analyzed problem. Rigorous analysis is needed to providc a solid framework for integrated networks, which are usually evaluated by means of approximations or simulation. In this appendix we review a formulation of the voice/data integration problem that is generic in nature and amenable to an optimization approach. Thus what has been a rather poorly analyzed problem can be subjected to rigorous analysis on which the ultimate design can be based.

We start from a routing problem first considered by Ephremides et al. (1980). A node with two outgoing link buffers was considered, and it was shown that a message should simply join the shortest queue. This model, despite its simplicity, proved to be rather difficult to analyze. It is not important to repeat the details here. It is sufficient to state that although the result was hardly surprising, a very intricate argument on the dynamic programming equation (DPE) was needed. Furthermore, some counterintuitive side results were obtained, including the observation that it was no longer necessary to require Poisson arrival statistics.

What is important is that this problem formulation is one of many related ones (see Nain and Ross 1986; Rosberg et al. 1982; Bhattacharya et al. 1987; and Hajek 1984). These formulations are slightly more complicated, but they share some fundamental characteristics. These characteristics, in fact, extend beyond the confines of the routing problem into the areas of priority assignment, resource allocation, and flow control. They are all Markovian decision process (MDP) problems. This appendix describes a fairly general MDP problem that includes the dynamic routing problem as a special case. In fact, it includes almost all of the queueing control problems that have been studied in connection with communication network issues. We then outline the solution methodologies that have been used. These are:

- the derivation of optimality conditions from the DPE that is associated with the corresponding MDP,

- the use of sample path or stochastic dominance arguments, and

- the reformulation of the MDP as a linear program (LP).

In fact, the last method seems to be quite powerful and possesses distinct advantages not shared by the other two. It was first used in Walrand (1984), and, more recently, the generality of it was demonstrated by Viniotis and Ephremides (1987, 1988a,b) and Viniotis (1988) who applied it to the analysis and control of the movable-boundary scheme. We must emphasize that the problems are quite complex, and only modest results can generally be obtained. Typically, these results do not provide the optimal control; they only characterize the structure of the optimal policy. However, the knowledge of the structure is often sufficient to permit close approximation of the actual policy by well-founded heuristics. In this appendix we discuss the application of these techniques to the voice/data integration problem.

Let us recall briefly what an MDP is (for details see e.g., Lippman 1973, 1975; Ross 1983; Tijms 1986). We need a state description of the process to be controlled. Let S be its state space. When in state $s \epsilon S$, a set A_s of admissible control actions is specified. When action $a \epsilon A_s$ is applied, a transition from state s to s' occurs that is governed by the probability distribution $q(s' | s,a)$, and which occurs after a random time τ that is exponentially distributed with the distribution denoted by $t(\tau | s,a,s')$. Clearly, q and t together describe the stochastic dynamics of the problem. Finally, this transition is accompanied by a cost penalty that we denote by $c(\tau | s,a,s')$.

We need to specify the notion of a control policy and the optimization criterion. Let us denote by $\xi_1, \xi_2, \cdots$ the state transitions that occur at instants $t_1, t_2, \cdots$. A policy π is a sequence of decision rules $\pi_1, \pi_2, \cdots$, where π_n determines the choice of action after the completion of the $(n-1)$-st transition. It can be viewed as a conditional distribution on the past history h^n of the process, which consists of the triplet (s_1, a_1, c_1) and the quadruples (τ_i, s_i, a_i, c_i) for $i = 1, \cdots, n-1$. If the action rule is nonrandom, we say that the policy is atomic. Finally, if π_n depends only on the current state, the policy is called stationary.

The optimization criterion that we generally prefer to consider is the long-run average, expected cost; namely, if we denote by $V(\pi, i, t)$ the expected cost incurred under policy π, with initial state i, until time t, we consider as the optimization criterion the function

$$V(\pi,i) \triangleq \lim inf_{t - \infty} \frac{V(\pi,i,t)}{t}.$$

However, for technical reasons that are well known to optimization specialists, optimality conditions are easier to establish if we consider, instead, the so-called α-discounted cost, i.e.,

$$V^{\alpha}(\pi,i) = \int_{t=0}^{\infty} e^{-\alpha t} V(\pi,i,t) dt.$$

The latter converges to the former as $\alpha \rightarrow 1$ under a variety of ergodicity-like conditions.

Dynamic Programming

Again for technical reasons that this time have to do with weaknesses of linear programming theory in infinite dimensions, we consider finite-horizon costs. These are defined in a similar fashion except that we let time extend only to t_n, the instant of the nth transition. If we denote the values of these cost functions by $V^{\alpha}(i)$ and $V(i)$ (and also $V_n^{\alpha}(i)$ and $V_n(i)$ for the finite-horizon cases) when π is chosen optimally, we are led to the dynamic programming equation (DPE), which assumes here the form (Bertsekas 1987):

$$V^{\alpha}(i) \;=\; \inf_{\alpha \epsilon A_i} \sum_{i'} [c(i,a,i') + \beta(i,a,i')V^{\alpha}(i')]\, q(i' \mid a,i)$$

or

$$V_{n+1}^{\alpha}(i) \;=\; \inf_{\alpha \epsilon A_i} \sum_{i'} [c(i,a,i') + \beta(i,a,i')V_n^{\alpha}(i')]\, q(i' \mid a,i),$$

where

$$\beta(s,a,s') \;\triangleq\; \int_0^{\infty} e^{-\alpha\tau} dt(\tau \mid s,a,s')$$

and

$$c(s,a,s') \;\triangleq\; \int_0^{\infty} c(\tau \mid s,a,s') dt(\tau \mid s,a,s')$$

are the discount factor and cost values per transition, respectively. In operator notation we can write

$$V_{n+1}^{\alpha}(i) \;=\; (TV_n^{\alpha})(i),$$

where T represents the DPE map from V_n^{α} to V_{n+1}^{α}.

The DPE is fundamentally important in the study of MDPs because the value function V^{α} usually has convenient properties of convexity, supermodularity, and other forms of monotonicity that readily lead to sufficient conditions for optimality. The difficulty with the analysis of the DPE is that the optimality conditions are heavily problem-dependent and often lead to explosively large numbers of cases to be verified separately. This is especially true for MDPs that arise from queueing models. For this reason, and because of additional difficulties that arise when the state is on the boundaries, it became evident that alternative methods of solution were needed.

Note also that the conditions of optimality usually supply only structural properties of the optimal policy. To determine the optimal policy exactly is a much more difficult task. Often it is necessary to settle on the related goal of determining the optimal value of the cost, $V^{\alpha}(i)$.* This can be done by the so-called value-iteration method by which almost any function f, under very general conditions, converges to $V^{\alpha}(i)$ as the dynamic programming operator acts on it, i.e.,

$$V^{\alpha}(i) \;=\; \lim_{n \to \infty} T^n f.$$

Sample-Path or Stochastic Dominance

One alternative method that has received attention recently, and which has produced successful results in problems of queueing control (akin to the routing problem), is a probabilistic method called sample-path or stochastic dominance (Walrand 1984). This method bypasses completely the need to deal with the value function. Instead it focuses directly on seeking the optimal policy. Let G be the class of admissible policies. If we suspect that the optimal policy π has a property p and wish to

*With this information, it is possible to determine how close a suboptimal policy, e.g., one obtained by heuristics, comes to the optimal.

prove that it actually does have that property, we can proceed as follows. Let S be a proper subset of G, to which we know the optimal policy belongs. We consider a subset of policies $S_p \subset S$, all elements of which have the property p. If π is not in the set S_p, we attempt to construct a policy $\tilde{\pi} \epsilon G - S_p$ that performs better than π. If we succeed, we have established a contradiction, and thus we must conclude that $\pi \epsilon S_p$. In constructing $\tilde{\pi}$ we often must carefully reorganize the underlying probability space to align the sample paths properly, so that the comparison of the two policies can be made for every sample path. This procedure is full of risks, and extreme care is required to avoid faulty arguments. Note, also, that to apply this method usefully, we must have correctly "guessed" the properties of the optimal policy. Thus, at best, it is a method of reorganizing our conclusions, rather than a method that leads us to the right conclusions.

Linear Programming

The third method is the linear programming (LP) method. It was first applied to the study of a specific queueing control problem by Rosberg et al. (1982), and recently was extended broadly by Viniotis (1988) for application to the voice/data integration problem. Very plainly, almost any queueing control problem that can be formulated as an MDP (therefore the problem of dynamic routing, as well) can be converted to an equivalent linear program. The advantages of this conversion are that it is less dependent on the specific features of the problem than dynamic programming, it leads to a complete characterization of the solution, and it sometimes leads to successful study of semi-Markovian decision problems as well. Furthermore, it facilitates the characterization of optimal solution properties. In Appendix B we demonstrate how MDPs can be converted to equivalent linear programs under very mild conditions that are usually satisfied by dynamic routing and other queueing control problems.

A Final Remark

We have discussed three possible approaches to the solution of MDPs. Whether to choose the dynamic programming approach, stochastic dominance tools, or linear programming, depends on the specific problem that is being addressed and on the (as yet undeveloped) intuition that we hope to develop in the course of this study.

REFERENCES

Bertsekas, D. (1987), *Dynamic Programming: Deterministic and Stochastic Models* (Prentice-Hall, Reading, MA).

Bhattacharya, P.P., I. Viniotis, and A. Ephremides (1987), "A (Not Very) Simple Routing Problem," *Proc. Twenty-Fifth Ann. Allerton Conf. Commun., Control, Comp.*, pp. 998-1006.

Bhattacharya, P.P. and A. Ephremides (1988), "Optimal Scheduling of the Transmission of Messages with Strict Deadlines," *Proc. 22nd Conf. Info. Sciences Syst. (CISS)*, Princeton University, pp. 623-628.

Bhattacharya, P.P. and A. Ephremides (1989), "Optimal Scheduling with Strict Deadlines," *IEEE Trans. Auto. Control* **34,** 721-728.

Ephremides, A., P. Varaiya, and J. Walrand (1980), "A Simple Dynamic Routing Problem," *IEEE Trans. Auto. Control* **AC-25**, 690-693.

Hajek, B. (1984), "Optimal Control of Two Interacting Servicc Stations," *IEEE Trans. Auto. Control* **AC-29**, 491-499.

Lippman, S.A. (1973), "Semi-Markov Decision Processes with Unbounded Rewards," *Management Sci.*, **19**, 717-731.

Lippman, S.A. (1975), "Applying a New Device in the Optimization of Exponential Queueing Systems," *Operations Res.* **23**, 687-710.

Nain, P. and K.W. Ross (1986) "Optimal Priority Assignment with Hard Constraint," *IEEE Trans. Auto. Control* **AC-31**, 883-888.

Rosberg, Z., P. Varaiya, and J.C. Walrand (1982), "Optimal Control of Service in Tandem Queues," *IEEE Trans. Auto. Control* **AC-27**, 600-610.

Ross, S.M. (1983), *Introduction to Stochastic Dynamic Programming* (Academic Press, New York).

Tijms, H.C. (1986), *Stochastic Modelling and Analysis: A Computational Approach*, (John Wiley and Sons, New York, NY).

Viniotis, I. and A. Ephremides (1987), "Optimal Switching of Voice and Data at a Network Node," *Proc. 28th Conference on Decision and Control (CDC)*, pp. 1504-1507.

Viniotis, I. (1988), "Optimal Control of Integrated Communication Systems Via Linear Programming Techniques," Ph.D. dissertation, University of Maryland.

Viniotis, I. and A. Ephremides (1988a), "On the Optimal Dynamic Switching of Voice and Data in Communication Networks," *Proc. Computer Networking Symp.*, pp. 8-16.

Viniotis, I. and A. Ephremides (1988b), "Linear Programming as a Technique for Optimization of Queueing Systems," *Proc. 27th Conf. Decision Control*, pp. 652-656.

Walrand, J. (1984), "A Note on Optimal Control of a Queueing System with Two Heterogeneous Servers," *Syst. Control Lett.* **4**, 131-134.

Appendix B

FORMULATION OF THE MDP AS A LINEAR PROGRAM

In Appendix A we note that almost any queueing control problem that can be formulated as an MDP can be converted to an equivalent linear program (LP). In this appendix we show how this conversion can be accomplished.

Let us concentrate on an MDP under a finite-horizon, discounted-cost formulation. The reason that we cannot work directly with infinite horizons is the possibility of so-called duality gaps in linear programming theory with infinite dimensional variables. We consider a queueing model with state dynamics given by a general equation of the form

$$x_{k+1} = x_k + \xi_{k+1} z_{k+1}.$$

Here, x_k (a vector) denotes the *state at* t_k *(the instant of the kth transition)*, ξ_k *(a matrix)* represents that transition, and z_k (a vector) represents the control action at that transition. For example, in Appendix A we mention a simple routing model in which an arriving message joins one of two output queues at a node. In this case, x_k would be a two-element vector whose elements are the two queue sizes. The transition matrix ξ_k can represent an arrival or a departure as an increment of the state. The control z_n is conveniently defined to enable ($z_k = 1$) or disable ($z_k = 0$) a transition. For example, sending an arriving message to the first queue would be represented by the first component of a two-dimensional vector z_k being equal to 1 and the second component of it being 0. (More generally, a probabilistic control could be also implemented under which the elements of the vector are the probabilities that the particular queues are chosen.) Typically, z_k would be chosen as a function of the state, with the goal of optimizing some performance criterion. Indeed, a broad variety of queueing control problems (in fact, the vast majority of those that have been considered in connection with communication network problems) can be represented in this manner.

The crucial aspect of this state equation is the linear dependence on the controls. Note also that the cost function is usually linear in the state (since the usual cost criterion is the expected delay, which is coupled to the queue sizes, and hence the state, by Little's result*). Consequently, the cost is linear in the controls. Selecting an optimal policy amounts to minimizing the cost by choosing the control actions. The minimization is constrained since the state equation must be satisfied, namely,

$$x_{k+1} = (x_k + \xi_k z_k) \in S.$$

*Little's result states that the average number of customers in a queueing system is equal to the average arrival rate times the average time a customer spends in the system (see e.g., Kleinrock 1975).

To reflect physical constraints, the states must be nonnegative ($x_k \geq 0$) and must not exceed the available buffer size ($x_k \leq B$). Thus the constraints are also linear in the controls. In short, this is an outline of the conversion to a linear program.

There are, however, points that require attention. First, the variables in a linear program take values in a continuum, e.g., $z_k \epsilon [0,1]$ or $z_k \epsilon R^n$. In contrast, in an unconstrained MDP problem the controls are integer-valued, i.e., $z_k \epsilon \{0,1\}$ or $\{0,1,2, \cdots ,N\}$. Second, in the MDP the z_n's are random, in general, and depend on the past history.

The first problem is taken care of in one of two ways: by construction, or by use of a property of the constraint matrix of the linear program called unimodularity. The construction method simply involves using a (noninteger) z^* that solves the problem whose restriction (integer-valued) z satisfies the MDP optimality conditions. The use of unimodularity involves a theorem that states that if in the linear program defined by

$$\min \; c'y \text{ subject to } Ay \leq b, \; y \geq 0$$

(where c' denotes the transpose of a constant column vector) A is totally unimodular (i.e., every subdeterminant of A is $+1$, -1, or 0), then the set of extreme points of the polyhedron $Ay \leq b$, $y \geq 0$ (to which, we know, the optimal y must belong) consists of integer-valued vectors. In many queueing problems of interest, including incidentally the dynamic routing problem, A is indeed totally unimodular.

The second problem is easily taken care of by thinking of the z_k's as functions from the sample space Ω to the action space. Thus the cost criterion can be written as a functional on the underlying probability space.

Let $z_k(\omega_k)$ represent the control action at the kth transition, with ω_k denoting the random "history" until the kth transition. We have

$$x_{k+1}(\omega_{k+1}) = x_k(\omega_k) + z_{k+1}(\omega_{k+1})\xi_{k+1}(\omega_{k+1}).$$

Let $\hat{x} = \{x_0, x_1, \cdots, x_n\}$, with $x_0 = x$, represent the trajectory of the state. A trajectory is feasible if $x_k(\omega_k) \epsilon S$, $\forall k$. Let Z be the set of admissible controls, and let Z_I be the subset of Z with integer valued z_k's. The β-discounted, n-step, expected cost under policy z and initial condition x is given by

$$J_n^\beta(x,z) = E_x \sum_{k=0}^{n-1} \beta^k(z_k),$$

where

$$L(z_k) = c'x_k + d'z_k$$

(where c' and d' denote transposed constant column vectors). This is a cost function that is adequately general. For example, in a pure resource-allocation problem without blocking or rejection of messages we have $d = 0$, whereas in pure blocking problems we take $c = 0$. The state equation, after repeated iteration, yields

$$x_k(\omega_k) = x + \sum_{j=1}^{k} z_j(\omega_j)\xi_j(\omega_j), \qquad k > 0.$$

Therefore,

$$J_n^\beta(x,z) = E_x \sum_{k=0}^{n-1} \beta^k \{c'x + c' \sum_{j=1}^{k} z_j \xi_j + d'z_k\}$$

$$= \frac{1-\beta^n}{1-\beta} c'x + E_x \sum_{k=1}^{n} \beta^k \{ \sum_{j=1}^{k} c'z_j \xi_j + d'z_k\}.$$

But

$$E_x(z_k) = \sum_{\omega_k} z_k Pr(\omega_k).$$

Hence

$$J_n^\beta(x,z) = \frac{1-\beta^n}{1-\beta} c'x + \sum_{k=1}^{n} \sum_{\omega_k} \gamma_k(\omega_k) z_k(\omega_k),$$

where $\gamma_k(\omega_k)$ is a known function that depends on $p(\omega_k)$, c, ξ_k, and β^k. Consequently, the MDP is equivalent to

$$\min_{z_k} \sum_{k=1}^{n} \sum_{\omega_k} \gamma_k(\omega_k) z_k(\omega_k)$$

subject to

$$\left[x + \sum_{j=1}^{k} z_j(\omega_j)\xi_j(\omega_j) \right] \epsilon \ S.$$

We recognize this last problem to be a classic linear program. Note that the initial condition plays the role of a parameter, the sensitivity with respect to which can be studied by the well-developed theory of sensitivity analysis of linear programming.

If we now let

$$V_n^\beta(x) \triangleq \min_{z \epsilon Z_I} \{J_n^\beta(x,z);\ \hat{x} \epsilon S\}$$

be the value function of the integer programming problem and

$$W_n^\beta(x) = \min_{z \epsilon Z} \{J_n^\beta(x,z);\ \hat{x} \epsilon S\}$$

be the value function of the linear program, we conclude that if the LP admits integer-valued solution, the two problems coincide and $V_n^\beta(x)$ is the restriction of $W_n^\beta(x)$. Thus all properties of W (convexity, supermodularity, etc.) are inherited by V. In particular, $V_n^\beta(x)$ is indeed an increasing, convex, supermodular, piecewise linear function of x.

In conclusion, we see that the MDP is converted to an equivalent LP under very mild conditions that are usually satisfied by dynamic routing and other queueing control problems.

REFERENCE

Kleinrock, L. (1975), *Queueing Systems, Vol. I: Theory* (John Wiley and Sons, New York, NY.

Appendix C

MDP MODEL FOR OPTIMIZATION OF VOICE/DATA INTEGRATION AT A SINGLE NODE

This appendix addresses the problem of the optimization of the movable-boundary scheme at a single node of an integrated network. We formulate the control problem as a Markovian decision process (MDP) problem, and we show how linear programming (LP) can be used to obtain the structure of the optimal policy. Extension of this model to multinode networks is a difficult problem. We discuss briefly the case of tandem networks for which some results have been obtained. The material in this appendix is based primarily on the recent work of Viniotis and Ephremides (Viniotis and Ephremides 1987, 1988a,b; Viniotis 1988).

We assume that a single link of bandwidth C serves voice and data that arrive at the node.* We use the standard terminology of "voice call" and "data message" to describe these two types of traffic; we sometimes use the simpler terminology of "call" and "message" since there is no ambiguity. The available bandwidth of the link is divided into N units of equal bandwidth C_o. C_o is the capacity required to serve a voice call. For simplicity, we assume that $C_o = 1$.

At any time t, $t \in [0, \infty]$, the link may serve i voice calls, $0 \leq i \leq N$, which consume a total of i units of the available bandwidth; the remaining capacity is allocated to data message traffic. In the present discussion, we assume that the entire bandwidth of $N - i$ frequency slots is allocated to only one data message. Messages that do not receive immediate service when they arrive are placed in a buffer, unlike voice calls, which are blocked (lost) if they cannot be serviced immediately. Flow-control mechanisms are needed to handle the possibility of overflowing a finite buffer. The assumption of very large (infinite) buffer capacity simplifies the flow-control aspect. In this appendix, we assume that the buffer capacity is infinite.

Note that the model considered in this appendix is somewhat different from that described in Section 7. Here we are considering a continuous-time frequency-division multiplexed system. In this system, each voice call is given a bandwidth allocation, and a single data message is given the remaining bandwidth. Since the number of calls varies with time, the bandwidth allocated to a message is time-varying and it may actually vary during the transmission of one particular message. Since it is assumed here that time is continuous, the time-varying rate of service does not result in gaps between messages as it would in a time-slotted system. Thus, such a model makes the analysis easier than that for time-division systems, such as those considered in Section 7. We believe that the optimal controls and resulting performance obtained for the continuous-time case provide a good indication of the performance of time-division systems as well, although they will not necessarily provide exactly optimal performance.

*When more than one link is connected to the node, the situation is different; data buffers may or may not be common to all links

When calls arrive, they are either given a bandwidth unit immediately or blocked (by choice or because there is no room for them). The decision to block a call may be dictated by the unavailability of resources at another node of the call path in the network, or it may be dictated by the need to service messages that are incurring long delays at a higher rate of service. When a call is accepted, the service rate for the message (if one is being served) is instantaneously decreased by one unit of bandwidth (C_o).

We adopt the following assumptions for the statistics of the arrival and service processes. Voice call arrivals form a Poisson stream of rate λ_c. Their corresponding holding times form a sequence of independent, identically distributed, exponential random variables with rate μ_c. Similarly, data messages are described by the parameters λ_m and μ_m. All processes are statistically independent from each other.

Under these assumptions, the number of messages in the system (those in queue plus the one in service) and t e number of calls being served, constitute a suitable state description of the system.*

We can now formulate the control problem in more precise terms, by using the theory of MDPs.

Control Problem Formulation

We wish to control the operation of the link (at transition instants) in the following way. When a message arrives to find one or more messages already in the system, it is placed in a queue.† When the arriving message finds no other messages in the queue and if i calls are already in service, it is immediately given $N - i$ bandwidth units, and service commences. Here we understand that when $i = N$, the message is placed in the queue.

When a call arrives to find $i = N$ calls already in service, it is immediately blocked (and lost), since no bandwidth is available. However, when it finds $i < N$ calls in service, a decision is made whether to accept it or not. The decision need not be deterministic and might be influenced by the availability of resources in other nodes of the network. The controller at a given node has to choose its actions so that a certain performance criterion (to be precisely defined later) is optimized. This is a dynamic form of the movable-boundary scheme under which the position of the boundary is chosen based on the instantaneous state of the node and traffic statistics, and possibly on the state at other nodes in the network as well.

When a data message completes service, another message from the queue (if nonempty) is immediately forwarded for service. Finally, when a call completes its service, the service rate for the message being served is instantaneously increased by one unit.

We see from the above considerations that calls are given "priority" over messages in the sense that the arrival (and acceptance) of a call results in the lowering of the message service rate by one unit. Also, calls are given priority in the sense that each call that is accepted maintains the use of a

*Clearly, some of these assumptions are not valid in a general multinode network, where the input at a particular node is the output of other node(s), possibly combined with exogenous new arrivals at the node. Few results are available for any but the simplest configurations, i.e., single nodes or simple tandems.

†When the queue has finite capacity, overflow may occur. This message is lost and must be retransmitted. Since retransmissions degrade system performance overflow has to be included somehow in the system performance measure. When the capacity of the queue is infinite, as we assume here, buffer overflow cannot occur.

dedicated unit of bandwidth C_o for its duration and is neither forced to wait nor can it be disrupted by data traffic. However, the possibility exists that a call will not be accepted so that data can be served more efficiently; the priority rule is not absolute.

We further assume that there is no interaction from other nodes in the sense that the decisions made at a particular node are not influenced by decisions at other nodes.

To formulate the optimal control problem as a Markovian decision process problem, we have to specify four elements:

- the state of the system,
- the controls,
- the (controlled) transitions among the states, and
- the cost functional (i.e., the cost incurred by the controls at each transition).

For any time $t \in [0, \infty]$, the vector $s_t \triangleq (i_t, j_t)$ is a suitable state of the system, which consists of a single node. Here, i_t denotes the number of calls being served at time t, and j_t denotes the total number of messages at the node (in the buffer and in service) at time t. Then $s_t \in S \triangleq \{0, 1, \cdots, N\} \times \{0, 1, 2, \cdots\}$, where S is the state space of the system.

Transitions among the states are described by the following operators (that map S into S):

call arrival: $A_c(i,j) = (i \oplus 1, j)$

message arrival: $A_m(i,j) = (i, j + 1)$

call completion: $D_c(i,j) = (i \ominus 1, j)$

message completion: $D_m(i,j) = (i, j \ominus 1)$

where we have used the notation $l \oplus 1 = \min\{N, l + 1\}$ and $l \ominus 1 = \max\{0, l - 1\}$.

We exercise control on call arrivals only. (All data messages are accepted and stored in an infinite buffer.) The decision to block or accept an incoming call need not be deterministic; it may depend on the history of the system up to t (i.e., state transitions and decisions). In general, the class of admissible controls is allowed to contain any nonanticipative rules. A policy is just a rule that prescribes actions at call arrival instants. Let t be such an instant. We denote the probability of rejecting an incoming call by z_t. The action of blocking a call (with probability 1) thus corresponds to setting $z_t = 1$. In general, we let $z_t \in [0, 1]$. Thus the action space is $A_t = [0, 1]$ A $t \in [0, \infty]$, such that t is a call arrival instant.

Under the statistical assumptions adopted before, s_t is a two-dimensional continuous-time Markov chain.

Notice that the transition rate "out" of a state varies from a minimum of $\lambda_m + \lambda_c$ to a maximum of $\lambda_m + \lambda_c + N$:max$\{\mu_m, \mu_c\}$. It is useful (when "discretizing" the continuous-time chain) to consider an imbedded chain for which the total rate out of a state does not depend on the state. Such a uniform discretization leads to a single discrete-time dynamic programming equation that holds true for any state, whether on the boundary or not. Properties of the optimal cost function are, therefore, easier to prove. Moreover, the discount factor (for discounted cost criteria) is constant, independent of the state. Thus, for "uniformization" purposes (Lippman 1975; Tijms 1986), we conveniently define the "total event rate" r* as

$$r = \lambda_c + \lambda_m + N\mu_m + N\mu_c.$$

Notice that r is greater than the maximum rate out of a state. Let $x \triangleq (i,j)$ be an arbitary state; let $I(A)$ denote the indicator function of the event A. For each control value z, define the transition probability function $p(\cdot \mid \cdot, z)$ by

$$\begin{aligned} p(y \mid x,z) \cdot r = \; & (1 - z)\lambda_c I(y = A_c x) \\ & + z\lambda_c I(y = x) \\ & + \lambda_m I(y = A_m x) \\ & + i\mu_c I(y = D_c x) \\ & + (N - i)\mu_m I(y = D_m x) \\ & + (\mu_m i + (N - i)\mu_c) I(y = x). \end{aligned}$$

The last term represents the dummy transition rate (i.e., transitions from a state into itself that simply make the total rate out of a state equal to r).

To complete the description oi the MDP, a performance criterion (cost function) should be specified. Since two different types of users are supported, the cost function must reflect the requirements of both. In general, cost criteria reflect performance requirements that are user-oriented or network-oriented. However, user-oriented requirements must eventually be mapped to network-oriented performance parameters thai are more applicable to the design and operation of the network.

Typically, probability of blocking and average delay are the most useful performance parameters for voice and data respectively.† The most frequently used criterion has been that of minimizing delay while keeping probability of blocking below certain thresholds.

*Any rate larger than $\lambda_m + \lambda_c + N\max\{\mu_m, \mu_c\}$ suffices for uniformization purposes; this particular choice of r will make the ω_k's described earlier identically distributed.

†In systems that use speech interpolation to increase the number of voice users, voice clipping is of primary interest as well, as discussed in Section 6.

This criterion leads to constrained MDP problems, a class of problems that have randomized optimal policies. In other words, when in state x, the policy takes an action a with some probability $p(x,a)$ that is not necessarily equal to 0 or 1. Such optimal probabilities are difficult to compute, even in simple optimal control problems. In a study of the movable-boundary scheme, Maglaris and Schwartz (1982) had to make an oversimplifying assumption to reduce the two-dimensional Markov chain to a set of two independent chains. Under this simplified model, deterministic policies yield the optimal solution.

Alternatively, a different criterion can be used, namely that of minimizing the weighted sum of the two performance parameters—delay and blocking probability. Such an approach is attractive for two reasons. First, the control problem is unconstrained, and therefore deterministic policies exist. Such policies are easier to characterize. Second, the weighting factor α is very similar to a Lagrange multiplier, and therefore this criterion can serve as an intermediate step for studying the constrained criterion problem. The cost functional we consider is, therefore,

$$EW + \alpha P_b,$$

where EW is the average delay for data messages and P_b is the probability that a voice call is blocked.

This criterion yields a continuous-time, infinite-horizon, undiscounted cost functional. However, it is much simpler to work with a discrete-time, finite-horizon, discounted cost functional* first, and then obtain the results for the original criterion, by considering appropriate limits.

In summary, our goal is to minimize (by blocking or accepting an arriving call) the weighted sum of the discounted delay for interactive data and blocking probability for calls. This problem is formulated as follows.

Let (i_t, j_t) be the system state at time t and (i,j) be the state at $t = 0$. Note that j_t is a measure of delay for data (since delay is directly proportional to the number of equal-size messages in the system), and z_i is a measure of blocking probability. We therefore let the instantaneous cost be given by $j_t + \alpha z_t$, where α is a weighting factor. In general $0 < \alpha < \infty$. For $\alpha = 0$ we have the trivial case where blocking an incoming call is of negligible importance. At the other extreme, for $\alpha = \infty$, blocking an incoming call is of paramount importance. The optimal policy for these limiting cases is simply always block or always accept, respectively.

Linear Programming Formulation

We now show how the optimization problem can be converted to a linear programming problem in the spirit of our earlier discussion.

We need to establish some notation and terminology first. We consider here the case in which the time horizon n is fixed and finite; k denotes the (discrete) time index.

*For discounted cost criteria, the dynamic programming (DP) operator is a contraction mapping; therefore, existence and uniqueness of solutions to the DP equation are guaranteed. On the other hand, it is not known a priori whether the same is true for undiscounted cost criteria.

Definition: A sample path ω^k is a sequence of events:

$$\omega^k \triangleq \{\omega_1, \omega_2, \cdots, \omega_k\}, \quad \text{where} \quad \omega_k \in \{A_c, A_m, D_c, D_m\} \;\; 1 \leq k \leq n.$$

A_m, A_c, D_m, and D_c are the transition operators corresponding to arrivals and departures of data messages and voice calls as defined earlier. Recall that uniformization introduces extra transitions. To be complete, we should allow in the sample path the event $\omega_k = D$, where $D(i,j) = (i,j)$; this is dummy transition for which there are no arrivals and no departures. However, since we have $\xi_k(\omega_k) = (0,0)$, the analysis (cost function expression and total unimodularity of the constraint matrix) is not affected. For this reason, we will ignore dummy transitions. For any k, let Ω^k denote the set of all events ω^k.

Definition: A transition ξ_k (at time k) is defined as the two-dimensional vector

$$\xi_k(\omega_k) = \begin{cases} (1,0) & \omega_k = A_c \\ (0,1) & \omega_k = A_m \\ (-1,0) & \omega_k = D_c \\ (0,-1) & \omega_k = D_m \\ (0,0) & \omega_k = D \end{cases} \tag{C1}$$

In other words, ξ_k denotes the change in state incurred by the kth transition.

Definition: A control policy $z \triangleq (z_1, z_2, \cdots, z_n)$ is defined as

$$z_k(\omega^k) = \begin{cases} 0 & \omega_k = A_m, \; D \\ \in [0,1] & \omega_k = A_c, \; D_c, \; D_m \end{cases} . \tag{C2}$$

Note that even though the control is allowed to depend on the entire "history" ω^k it actually depends only on the most recent state ω^k, i.e., it is Markovian. This is not a restriction, since it is known that the optimal control is indeed Markovian.*

Definition: A state trajectory $\hat{x} \triangleq (x_1, x_2, \cdots, x_n)$ is defined as

$$x_0 = x \triangleq (i,j)$$

$$x_k(\omega^k) = x_{k-1}(\omega^{k-1} + (1 - z_k(\omega^k)) \, \xi_k(\omega^k). \tag{C3}$$

We see from Eq. (C2) that when the transition is a call arrival, we allow $z_k = 0$ or 1, and thus the transition is possibly disabled. However, when we have $\omega_k \neq A_c$, the control z_k should be set to

*In fact, it is the feature that makes us believe that the LP technique might be useful for semi-Markovian problems as well. Problems arising in integrated networks that use TDMA multiplexing techniques are of that nature.

0 to allow the transition, since we are assuming that all data messages are accepted and stored in the buffer; no control action is taken on data messages. The uniformization procedure, however, introduces certain "dummy" departures of calls and messages from states with zero calls or messages. Such (fictitious) transitions would result in states with negative components, and they should of course be disabled. Therefore, "fictitious" controls at departures should be introduced.*

The cost incurred by a policy z is denoted as $J_n(z,i,j)$. The elements of ξ_k, x_k are denoted as ξ_{k1}, ξ_{k2}, and x_{k1}, x_{k2}. The first element refers to calls; the second one refers to data messages.

Following the procedure described earlier, we may rewrite the cost incurred by z as

$$J_n(z,i,j) \; = \; (1 + \cdots + \beta^{n-1})j + \sum_{k=1}^{n} \sum_{\omega^k \epsilon \Omega^k} \gamma_k(\omega^k)\, z_k(\omega^k), \tag{C4}$$

where the coefficients $\gamma_k(\omega^k)$ are given by

$$\gamma_k(\omega^k) \; = \; \begin{cases} 0 & \omega_k = D_c \\ Pr(\omega^k)(\beta^k + \cdots + \beta^{n-1})\xi_{k2}(\omega_k) & \omega_k = A_m, D_m \; . \\ Pr(\omega^k)(\beta^k + \cdots + \beta^{n-1})\alpha & \omega_k = A_c \end{cases} \tag{C5}$$

The above selection of $\gamma_k(\omega^k)$ does not incorporate departure controls into the cost; therefore blocking probability is correctly expressed. The crucial point is that $J_n(z,i,j)$ depends linearly on $z_k(\omega^k)$. Taking liberty with notation, let us write $z_k(\omega^k,z)$ to denote the dependence of the state on the control. Since the term $(1 + \cdots + \beta^{n-1})j$ does not depend on the policy in operation, the optimal policy is characterized by the following LP problem:

$$\min_{z} \sum_{k=1}^{n-1} \sum_{\omega^k \epsilon \Omega^k} \gamma_k(\omega^k) z_k(\omega^k) \tag{C6}$$

under the constraints

$$\begin{array}{rl} 0 \leq z_k(\omega^k) \leq 1 & \omega_k = A_c, D_c, D_m \\ z_k(\omega^k) = 0 & \omega_k = A_m, D \end{array} \tag{C7}$$

(a) (nonnegative states)

$$x_k(\omega^k,z) \; = \; x + \sum_{j=1}^{k} (1 - z_j(\omega^j))\xi_j(w^j) \geq 0 \tag{C8}$$

(b) (bounded number of calls)

$$x_{k1}(\omega^k,z) = i + \sum_{j=1}^{k} (1 - z_j(\omega^j))\xi_{j1}(\omega^j) \leq N. \tag{C9}$$

*This requires extra caution, since the instantaneous cost includes z_k. The cost of disabling departure should not be included in the call blocking probability. As we shall see, proper selection of the cost coefficients $\gamma_k(\omega^k)$ takes care of this problem.

The set of constraints, Eqs. (C8) (C9), for $1 \leq k \leq n, \omega^k \epsilon \Omega^k$, can be written in compact form as

$$AZ \leq Fb + b_0. \tag{C10}$$

The vector b is defined as $b \triangleq [i\ j]$. The matrix F is

$$F = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -1 \end{pmatrix},$$

where 0, 1 and -1 denote strings of all 0, 1, and -1 respectively. Also,

$$b_0 = \begin{pmatrix} 0 \\ 0 \\ N \end{pmatrix}.$$

The LP formulation of the optimal control problem is now complete; the theory described earlier can be applied to yield the properties of the optimal cost function.

Structure of the Optimal Policy

Although we are primarily interested in the infinite-horizon, average-cost criterion, we perform an intermediate step to obtain the structure of the optimal policy for discounted and finite-horizon criteria.

One can show (Viniotis 1988) that the cost function $J - n(i,j)$ is a convex and supermodular function of the state (i,j), and that by taking the limit as $n \rightarrow \infty$, we have that the infinite-horizon cost function $J(i,j)$ is a convex and supermodular function ot the state (i,j).

Recall that both $J_n(i,j)$ and $J(i,j)$ refer to discounted cost criteria.

Let $h_{ij}(k) \triangleq J_k(1,j)$. Since the function $J_k(\cdot,\cdot)$ is monotone in k and bounded, the limit

$$h_{ij} \triangleq \lim_{k \rightarrow \infty} h_{ij}(k)$$

is well defined for any i,j. Note that the condition $h_{ij} > \alpha/\beta\lambda_c$ implies that blocking is optimal any time a call arrives to find the system at state (i,j).

For each j, define the function

$$i' \triangleq S(j) \triangleq \min\left\{0 \leq i \leq N : h_{ij} > \frac{\alpha}{\beta\lambda_c}\right\}.$$

In other words, i' is the smallest integer such that (i',j) is a blocking state. Note that (N,j) is a blocking state for any j. For finite-horizon criteria, the function $S_n(j)$ may be similarly defined.

Since the optimal cost function is convex in i, we have that

$$h_{ij} > \frac{\alpha}{\beta\lambda_c} \rightarrow h_{i+1,j} > \frac{\alpha}{\beta\lambda_c}.$$

In other words, state (i,j) is also blocking for any $i > i' = S(j)$. It is intuitively apparent that blocking a call decreases the waiting time portion of the cost for data messages, since they continue service at a higher rate. If with i calls already being served, it is optimal not to lower the service rate for messages, it should also be optimal not to lower it when more than i calls are already in the system. Intuition also suggests that state $(i,j + 1)$ be blocking, since with one more message, the waiting time portion of the cost (which "dictates" blocking at state (i,j)) can only become worse. This is exactly what supermodularity of the cost function suggests, since we have that

$$h_{ij} > \frac{\alpha}{\beta\lambda_c} \rightarrow h_{i,j+1} > \frac{\alpha}{\beta\lambda_c}.$$

Viniotis and Ephremides (1988a) show that the general form of the solution is a monotone switching curve, such as that shown in Fig. C1. The switching curve divides the state space into two regions, one in which calls are blocked and the other in which they are not. The same form characterizes the structure of the optimal policy for finite-horizon criteria as well. Of course, in this case the curve depends on the horizon n, since it is generally nonstationary (but still Markovian). Computations have shown that the switching curve is a monotone-decreasing function of n, i.e., $S_n(j) \geq S_{n+1}(j)$. However, this has not been established rigorously.

To summarize, it has been shown that the policy that minimizes the weighted sum of blocking probability and data delay is a "switching curve" policy.

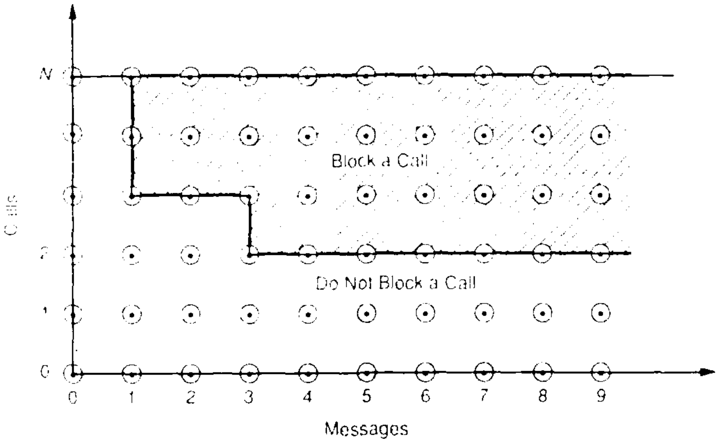

Fig. C1 — Form of the switching curve for optimization of the movable-boundary scheme

Implementation Issues

The switching-curve result just discussed is an existence result; it describes only the structure of the optimal policy. The exact dependence of the switching curve on model parameters is quite difficult to obtain in all related queueing-control problems.

Two possible courses of action can be followed for implementation. The first involves approximating the switching curve with simpler curves (suboptimal controls). The second involves approximation of the switching curve by using the value-iteration technique of dynamic programming. From computational experience, for all the related queueing models in the literature and for a wide range of model parameters, the value-iteration algorithm usually converges (for discounted costs) to the optimal policy in less than 50 iterations.

The Tandem Network

Consider a configuration of M nodes in tandem (Fig. C2) where each node in the network relays the traffic (both voice and data) it has received from the previous (upstream) node, as well as new traffic (called exogenous traffic) generated at that node. All traffic is to be transmitted "downstream" (i.e., to higher numbered nodes). We assume that all traffic exits at the last node (M) of the network.

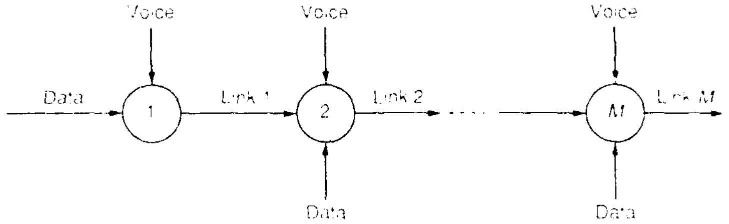

Fig. C2 — A tandem communication network for voice and data

Again the control parameter at our disposal is acceptance or blocking of incoming calls at the various nodes in the network. Decisions are centralized, since we require that the state of the network be known at all nodes. Note that acceptance of a call at one node implies a real-time commitment of resources at all downstream nodes. Decentralized controls, which use partial state information, have not been considered yet and will be an important subject of future research.

It is possible to use a Markovian model that describes the operation of the multiplexed channels. The differences between this model and the single-node model considered before are substantial. However, we can still formulate the optimal control problem as an MDP problem. The LP formulation is also valid, and the structure of the optimal policy has again been shown to be of the switching curve variety (Viniotis 1988).

Data Routing in Integrated Networks

The problem of data routing in integrated networks has received so far no attention in the literature. One reason is that data routing does not, of course, arise in single-node models. Another reason is that even in simple networks (optimal) routing is quite complicated. As a first step toward addressing this problem, we consider a configuration of M nodes in parallel; this model provides the designer with an extra control parameter, namely data routing.*

More specifically, we consider that the network consists of M nodes, each of which receives its own voice traffic. We have the option of accepting or blocking a call that arrives at node l, and routing a message to any one of the M nodes. The controller has to choose blocking and message routing actions to minimize the usual cost functional of weighted data delay and call blocking probability.

The queueing model for the routing decision is similar to that studied by Ephremides et al. (1980). The fundamental difference that complicates the present model is that service rate for messages is no longer constant but rather is a random variable.

Again it is possible to formulate the model for this network, together with the associated optimal control problem. A linear programming problem conversion can be obtained in a similar fashion, and the optimal policy can again be shown to have a switching-curve form.

REFERENCES

Ephremides, A., P. Varaiya, and J. Walrand (1980), "A Simple Dynamic Routing Problem," *IEEE Trans. Auto. Control* **AC-25**, 690-693.

Lippman, S.A. (1975), "Applying a New Device in the Optimization of Exponential Queueing Systems," *Operations Res.* **23**, 687-710.

Maglaris, B.S. and M. Schwartz (1982), "Optimal Fixed Frame Multiplexing in Integrated Line- and Packet-Switched Communication Networks," *IEEE Trans. Info. Theory* **IT-28**, 263-273.

Tijms, H.C. (1986), *Stochastic Modelling and Analysis: A Computational Approach* (John Wiley and Sons, New York, NY).

Viniotis, I. and A. Ephremides (1987), "Optimal Switching of Voice and Data at a Network Node," *Proc. 28th Conf. Decision Control (CDC)*, pp. 1504-1507.

Viniotis, I. (1988), Optimal Control of Integrated Communication Systems Via Linear Programming Techniques," Ph.D. dissertation, University of Maryland.

Viniotis, I. and A. Ephremides (1988a), "On the Optimal Dynamic Switching of Voice and Data in Communication Networks," *Proc. Computer Networking Symp.*, pp. 8-16.

Viniotis, I. and A. Ephremides (1988b), "Linear Programming as a Technique for Optimization of Queueing Systems," *Proc. 27th Conf. Decision Control*, pp. 652-656.

*When the data messages are routed to a queue in a probabilistic fashion, call blocking is the only control parameter.

Appendix D

A VOICE/DATA INTEGRATION SCHEME BASED ON THE 2400 b/s LPC SYSTEM

The movable-boundary scheme (combined with digital speech and data interpolation) is the most widely used approach for voice/data multiplexing. A totally different approach for the integration of voice and data has recently been developed by Kang (1985), who demonstrated that a low-rate data stream can be transmitted interspersed with 2400 b/s linear predictive coded (LPC) speech without degrading speech quality. This approach is based on the use of a single voice channel, rather than the multiple-voice-stream multiplexing approaches considered in Section 7, and is meant for the combining of data and voice from a single source to a single destination, rather than the multiplexing of voice and data from many users over a single link.

The approach is based on the observation that some of the bits in the 2400 b/s LPC bit-stream can be dropped without affecting speech quality. These bits could then be replaced by data. Like the TADI approach discussed in Section 7, normal gaps in speech can be replaced by data.* In addition, it has been demonstrated that unvoiced speech (consonants) requires fewer bits to encode than voiced speech (vowels). Four bits of data (which require the transmission of 8 bits when a rate 1/2 Hamming (8,4) block code is used) can be transmitted per unvoiced frame (of duration 22.5 ms). Approximately 40 to 50% of frames are unvoiced, depending on the particular talker's speaking pattern, and most of these unvoiced frames can be detected reliably and used for data. This results in a data rate between 71 and 88 b/s.

Kang (1985) estimated that approximately 75 b/s of data can be transmitted during most speech without loss of speech intelligibility. This low-rate data channel may be used to transmit data that can supplement the information transmitted by speech, e.g., numbers (which are easily forgotten or misunderstood when available only by speech) or visual aids such as simple hand-sketched drawings.† Furthermore, the combined voice/data mode is compatible with other 2400 b/s LPC systems that do not have this capability.

In this report we emphasize the multiplexing and switching issues that permit the flexible sharing of nodal and channel resources among a number of voice calls and data messages that may have different sources and destinations. However, approaches such as the one outlined in this appendix are useful in augmenting the communication capability that can be achieved between a single source-destination pair, and they may be used in conjunction with the hybrid switching techniques we have described.

REFERENCE

1. Kang, G.S. (1985), "Narrowband Integrated Voice Data System Based on the 2400-b/s LPC," NRL Report 8942.

*Here we are assuming that speech is a continuous one-way stream that is interrupted only by short gaps (a small fraction of a second) between talkspurts.

†Kang(1985) provided an example of a simple map showing enemy concentrations that requires 904 bits to describe. Such a figure requires 12 s to transmit at 75 b/s.